第一章 条件期望、条件方差和线性投影

内容提要

- 条件期望 1.1
- 简单迭代期望律 1.1
- 迭代期望律 1.2
- 条件方差 1.2
- 方差分解 1.4
- 线性投影 1.3

条件期望在计量经济学中扮演着重要的角色，古典线性回归模型的核心估计方法最小二乘（Least squares）本质上是在估计以解释变量 $\mathbf{x}$ 为条件被解释变量 y 的条件期望。而且，根据 $\mathbf{x}$ 提供的信息，关于 y 的最好（最小化均方误差）的预测也正是条件期望 $\mathbb{E}(\mathrm{y}|\mathbf{x})$。因此，本章将对条件期望的定义及其性质做一个较为完整的梳理。同时，本章还将讨论条件方差以及条件期望的最优线性近似（也就是：线性投影）的定义和性质。想深入了解本章内容可以进一步参考 Angrist et al. (2008); Hansen (2022a); Wooldridge (2010)。

1.1 条件期望

定义 1.1. 条件期望

令 $\mathbf{x}$ 为 K 维随机向量，y 为一元随机变量。以 $\mathbf{x}$ 为条件，y 的条件期望函数（Conditional Expectation Function, CEF）定义如下：

$$\mu(\mathbf{x}) = \mathbb{E}(\mathrm{y}|\mathbf{x}) = \begin{cases} \int_{\mathrm{y}} y \cdot f(y|\mathbf{x})dy, & \mathrm{y}\ \text{连续}, \\ \sum_{\mathrm{y}} y \cdot \mathbb{P}_{\mathrm{y}|\mathbf{x}}(y|\mathbf{x}), & \mathrm{y}\ \text{离散}, \end{cases}$$

其中，$f(y|\mathbf{x})$ 是以 $\mathbf{x}$ 为条件 $\mathrm{y} = y$ 的条件概率密度函数，$\mathbb{P}_{\mathrm{y}|\mathbf{x}}(y|\mathbf{x})$ 是以 $\mathbf{x}$ 为条件 $\mathrm{y} = y$ 的条件概率。 ♣

在定义1.1中，条件期望函数被记为 $\mu(\mathbf{x})$，这表明 CEF 是随机向量 $\mathbf{x}$ 的函数，与随机变量 y 的具体实现值无关。关于条件期望，我们有如下一些重要的结论：

定理 1.1. 简单迭代期望律（Simple Law of Iterated Expectations, SLIE）

若 $\mathbb{E}|\mathrm{y}| < \infty$，则对任意的随机向量 $\mathbf{x}$，我们有：

$$\mathbb{E}_{\mathbf{x}}[\mathbb{E}[\mathrm{y}|\mathbf{x}]] = \mathbb{E}[\mathrm{y}],$$

其中，符号 $\mathbb{E}_{\mathbf{x}}[\cdot]$ 表示相对于随机向量 $\mathbf{x}$ 求期望。 ♡

证明 假设 y 和 $\mathbf{x}$ 的联合概率密度函数为 $f(y, \boldsymbol{x})$，因为 $\mathbb{E}[\mathrm{y}|\mathbf{x}]$ 是 $\mathbf{x}$ 的函数，我们有：

$$\mathbb{E}[\mathbb{E}[\mathrm{y}|\mathbf{x}]] = \int_{\mathbb{R}^k} \mathbb{E}[\mathrm{y}|\boldsymbol{x}] f_{\mathbf{x}}(\boldsymbol{x}) d\boldsymbol{x}.$$

将 y 为连续随机变量情况下条件期望的定义1.1代入上式得到：

$$\int_{\mathbb{R}^k} \left(\int_{\mathbb{R}} y f(y|\boldsymbol{x}) dy \right) f_{\mathbf{x}}(\boldsymbol{x}) d\boldsymbol{x} = \int_{\mathbb{R}^k} \int_{\mathbb{R}} y f(y, \boldsymbol{x}) dy d\boldsymbol{x} = \int_{\mathbb{R}} y f(y) dy = \mathbb{E}[\mathrm{y}],$$

其中，第一个等号成立是因为 $f(y|\boldsymbol{x})f_{\mathbf{x}}(\boldsymbol{x}) = f(y,\boldsymbol{x})$；第二个等号成立是因为联合分布和边缘分布之间存在如下关系：$\int_{\mathbb{R}^k} f(y,\boldsymbol{x})d\boldsymbol{x} = f(y)$。 □

为了更好地理解定理 1.1，我们假设 $\mathbf{x}$ 是离散随机变量，则有：

$$\mathbb{E}[\mathbb{E}[\mathrm{y}|\mathbf{x}]] = \sum_{j=1}^{\infty} \mathbb{E}\left[\mathrm{y}|\mathbf{x}=\boldsymbol{x}_j\right]\mathbb{P}\left[\mathbf{x}=\boldsymbol{x}_j\right].$$

如果我们将求期望理解为求平均，那么条件均值的加权（以 $\mathbf{x}$ 取不同值的概率作为权重）平均是无条件均值。例如，令 $\mathbf{x}$ 代表性别，y 代表身高，

$$\mathbb{E}[\text{身高} \mid \text{性别}=\text{女}]\mathbb{P}\left[\text{性别}=\text{女}\right] + \mathbb{E}[\text{身高} \mid \text{性别}=\text{男}]\mathbb{P}\left[\text{性别}=\text{男}\right] = \mathbb{E}[\text{身高}].$$

所以，如果我们想得到一个人口样本的平均身高，可以直接求平均身高，也可以分别先求得女性和男性的平均身高再以各个性别所占的比例作为权重取加权平均。

定理 1.2. 迭代期望律（Law of Iterated Expectations, LIE）

若 $\mathbb{E}|\mathrm{y}| < \infty$，则对任意的随机向量 $\mathbf{x}_1$ 和 $\mathbf{x}_2$，我们有：

$$\mathbb{E}_{\mathbf{x}_1}[\mathbb{E}[\mathrm{y}|\mathbf{x}_1,\mathbf{x}_2]|\mathbf{x}_1] = \mathbb{E}[\mathrm{y}|\mathbf{x}_1].$$

♡

证明 为了简化证明，这里假设定理中所涉及的随机变量具有联合密度函数。注意到

$$f\left(y \mid \boldsymbol{x}_1,\boldsymbol{x}_2\right) f\left(\boldsymbol{x}_2 \mid \boldsymbol{x}_1\right) = \frac{f\left(y,\boldsymbol{x}_1,\boldsymbol{x}_2\right)}{f\left(\boldsymbol{x}_1,\boldsymbol{x}_2\right)}\frac{f\left(\boldsymbol{x}_1,\boldsymbol{x}_2\right)}{f\left(\boldsymbol{x}_1\right)} = f\left(y,\boldsymbol{x}_2 \mid \boldsymbol{x}_1\right), \tag{1.1}$$

以及，

$$\mathbb{E}[\mathrm{y}|\mathbf{x}_1=\boldsymbol{x}_1,\mathbf{x}_2=\boldsymbol{x}_2] = \int_{\mathbb{R}} y f(y \mid \boldsymbol{x}_1,\boldsymbol{x}_2)dy. \tag{1.2}$$

令随机向量 $\mathbf{x}_2$ 的维度为 k_2，于是，我们有：

$$\begin{aligned}
\mathbb{E}[\mathbb{E}[\mathrm{y}|\mathbf{x}_1,\mathbf{x}_2]|\mathbf{x}_1=\boldsymbol{x}_1] &= \int_{\mathbb{R}^{k_2}} \mathbb{E}[\mathrm{y}|\mathbf{x}_1=\boldsymbol{x}_1,\mathbf{x}_2=\boldsymbol{x}_2]f(\boldsymbol{x}_2 \mid \boldsymbol{x}_1)d\boldsymbol{x}_2 \\
&= \int_{\mathbb{R}^{k_2}} \left(\int_{\mathbb{R}} y f(y \mid \boldsymbol{x}_1,\boldsymbol{x}_2)dy\right) f(\boldsymbol{x}_2 \mid \boldsymbol{x}_1)d\boldsymbol{x}_2 \\
&= \int_{\mathbb{R}^{k_2}} \int_{\mathbb{R}} y f(y \mid \boldsymbol{x}_1,\boldsymbol{x}_2) f(\boldsymbol{x}_2 \mid \boldsymbol{x}_1)dyd\boldsymbol{x}_2 \\
&= \int_{\mathbb{R}^{k_2}} \int_{\mathbb{R}} y f\left(y,\boldsymbol{x}_2 \mid \boldsymbol{x}_1\right)dyd\boldsymbol{x}_2 \\
&= \mathbb{E}[\mathrm{y}|\mathbf{x}_1=\boldsymbol{x}_1],
\end{aligned}$$

其中，第二个等式用到了式 (1.2)；第四个等式用到了式 (1.1)。

由上式易知，$\mathbb{E}[\mathbb{E}[\mathrm{y}|\mathbf{x}_1,\mathbf{x}_2]|\mathbf{x}_1] = \mathbb{E}[\mathrm{y}|\mathbf{x}_1]$。 □

为了便于记住定理 1.2，我们可以把此性质总结为“小信息集占优”。此定理在时间序列分析中有广泛的应用，例如，令 $\mathscr{F}_t$ 表示 t 时刻所能获得的信息，我们有：若 $s < t$，$\mathbb{E}\{\mathbb{E}[\mathrm{y}_{t+h}|\mathscr{F}_t]|\mathscr{F}_s\} = \mathbb{E}[\mathrm{y}_{t+h}|\mathscr{F}_s]$。实际上，可以证明如下更一般形式的迭代期望律：令 $\mathbf{w}$ 为随机向量，$\mathbf{x}$ 是随机向量且满足 $\mathbf{x} = \mathbf{g}(\mathbf{w})$[1]。一般形式的 LIE 可表示为：

$$\mathbb{E}[\mathrm{y}|\mathbf{x}] = \mathbb{E}\left[\mathbb{E}(\mathrm{y}|\mathbf{w})|\mathbf{x}\right].$$

[1] 如果已知 $\mathbf{w}$ 的实现值，$\mathbf{x}$ 的实现值便是已知的。

定理 1.3. 条件化定理（Conditioning Theorem, CT）

若 $\mathbb{E}|\mathrm{y}| < \infty$，则，

$$\mathbb{E}[g(\mathbf{x}) \cdot \mathrm{y}|\mathbf{x}] = g(\mathbf{x}) \cdot \mathbb{E}[\mathrm{y}|\mathbf{x}].$$

若再假设 $\mathbb{E}|g(\mathbf{x})| < \infty$，则，

$$\mathbb{E}[g(\mathbf{x}) \cdot \mathrm{y}] = \mathbb{E}[g(\mathbf{x}) \cdot \mathbb{E}[\mathrm{y}|\mathbf{x}]].$$

♡

证明 给定 $\mathbf{x} = \boldsymbol{x}$，我们有：

$$\mathbb{E}[g(\mathbf{x}) \cdot \mathrm{y}|\mathbf{x} = \boldsymbol{x}] = \int_{\mathbb{R}} g(\boldsymbol{x}) y f(y|\boldsymbol{x}) dy = g(\boldsymbol{x}) \int_{\mathbb{R}} y f(y|\boldsymbol{x}) dy = g(\boldsymbol{x}) \mathbb{E}(\mathrm{y}|\mathbf{x} = \boldsymbol{x}).$$

因此，$\mathbb{E}[g(\mathbf{x}) \cdot \mathrm{y}|\mathbf{x}] = g(\mathbf{x}) \cdot \mathbb{E}[\mathrm{y}|\mathbf{x}]$。由定理 1.1，本定理中的第二个式子容易验证。 □

下面我们给出关于条件期望函数的两个重要性质：

性质 （CEF 分解性质）对任意的随机变量 y，有如下性质成立：

$$\mathrm{y} = \mathbb{E}[\mathrm{y}|\mathbf{x}] + \varepsilon, \tag{1.3}$$

其中，

(a) ε 均值独立于 $\mathbf{x}$，即，$\mathbb{E}[\varepsilon|\mathbf{x}] = 0$；

(b) ε 与 $\mathbf{x}$ 的任意函数都不相关。

证明 首先证明 (a)，

$$\mathbb{E}[\varepsilon|\mathbf{x}] = \mathbb{E}\{\mathrm{y} - \mathbb{E}[\mathrm{y}|\mathbf{x}]|\mathbf{x}\} = \mathbb{E}[\mathrm{y}|\mathbf{x}] - \mathbb{E}[\mathrm{y}|\mathbf{x}] = 0.$$

其次证明 (b)，假设 $h(\mathbf{x})$ 是 $\mathbf{x}$ 的任意函数，我们有：

$$\begin{aligned}
\mathrm{Cov}[\varepsilon, h(\mathbf{x})] &= \mathbb{E}[\varepsilon \cdot h(\mathbf{x})] - \mathbb{E}(\varepsilon)\mathbb{E}[h(\mathbf{x})] \\
&= \mathbb{E}\{\mathbb{E}[\varepsilon \cdot h(\mathbf{x})|\mathbf{x}]\} - 0 \cdot \mathbb{E}[h(\mathbf{x})] && \text{（这里使用了两次定理 1.1）} \\
&= \mathbb{E}\{h(\mathbf{x})\mathbb{E}[\varepsilon|\mathbf{x}]\} && \text{（这里使用了定理 1.3）} \\
&= 0. && \text{（这里使用了 (a) 的结论）}
\end{aligned}$$

□

性质 （CEF 预测性质）假设 $\mathbb{E}(\mathrm{y}^2) < \infty$，记 $\mu(\mathbf{x}) \equiv \mathbb{E}(\mathrm{y}|\mathbf{x})$，则 μ 是如下优化问题的解：

$$\min_{m \in \mathscr{M}} \mathbb{E}[(\mathrm{y} - m(\mathbf{x}))^2],$$

其中，$\mathscr{M}$ 是函数类，$m : \mathbb{R}^K \to \mathbb{R}$ 满足 $\mathbb{E}[m(\mathbf{x})^2] < \infty$。

证明 首先，我们对优化的目标函数进行分解：

$$\begin{aligned}
& \mathbb{E}[(\mathrm{y} - m(\mathbf{x}))^2] \\
= \ & \mathbb{E}[(\mathrm{y} - \mu(\mathbf{x}) + \mu(\mathbf{x}) - m(\mathbf{x}))^2] \\
= \ & \mathbb{E}[(\mathrm{y} - \mu(\mathbf{x}))^2] + \mathbb{E}[(\mu(\mathbf{x}) - m(\mathbf{x}))^2] + 2 \cdot \mathbb{E}\{[\mathrm{y} - \mu(\mathbf{x})] \cdot [\mu(\mathbf{x}) - m(\mathbf{x})]\}.
\end{aligned}$$

注意到，我们需要从函数类 $\mathscr{M}$ 中选择函数 m 使目标函数最小化，上述分解中的第一项与 m 的选择无关，第二项在 $m(\mathbf{x}) = \mu(\mathbf{x})$ 时取到最小值 0，我们只需证明第三项的取值与 m 的选择无关，实际上，

$$\mathbb{E}\{[\mathrm{y} - \mu(\mathbf{x})] \cdot [\mu(\mathbf{x}) - m(\mathbf{x})]\} = \mathbb{E}\{\varepsilon \cdot [\mu(\mathbf{x}) - m(\mathbf{x})]\} = 0,$$

其中，第二个等号成立是因为 $\mu(\mathbf{x})-m(\mathbf{x})$ 是 $\mathbf{x}$ 的函数，利用 CEF 分解性质 (b) 即可。

于是，目标函数的最小值在 $m(\mathbf{x})=\mu(\mathbf{x})$ 时取到，此性质得证。 □

对 $\mathbf{x}$ 任意的函数 $m(\mathbf{x})$，定义其预测 y 得到的均方误差（mean squared error, MSE）为

$$\mathrm{MSE}(\mathrm{y};m)\equiv\mathbb{E}[(\mathrm{y}-m(\mathbf{x}))^2],$$

CEF 的分解性质表明：对任意的随机变量 y 和随机向量 $\mathbf{x}$，y 总是可以分解为 $\mathbf{x}$ 可以解释的部分 $\mathbb{E}[\mathrm{y}|\mathbf{x}]$ 和与 $\mathbf{x}$ 不相关的部分 ε。对于 $\mathbf{x}$ 可以解释的部分 $\mathbb{E}[\mathrm{y}|\mathbf{x}]$，CEF 的预测性质指出它是关于 y 的最小化均方误差的预测。计量经济学的重要任务之一就是研究如何基于已有的信息对感兴趣的变量进行预测，因此，条件期望是计量经济学重要的研究对象。值得一提的是，传统的计量方法通常假设条件期望函数具有线性函数形式，但实际上 $\mathbb{E}[\mathrm{y}|\mathbf{x}]$ 可以是 $\mathbf{x}$ 的非线性函数。为此，计量经济学也发展出诸如参数非线性模型、非参数或半参数模型、神经网络模型等一系列非线性模型。另外，计量经济学还包括公司金融等微观实证研究使用的计量方法，特别是关于因果推断或政策评估的计量方法，这些方法关注的重点不是条件期望，而是一些因果效应参数的估计。

1.2 条件方差

定义 1.2. 条件方差

令 $\mathbf{x}$ 为随机向量，y 为一元随机变量。如果 $\mathbb{E}[\mathrm{y}^2]<\infty$，以 $\mathbf{x}=\boldsymbol{x}$ 为条件，y 的条件方差（Conditional Variance）定义如下：

$$\sigma^2(\boldsymbol{x})=\mathrm{Var}[\mathrm{y}|\mathbf{x}=\boldsymbol{x}]=\mathbb{E}\left[(\mathrm{y}-\mathbb{E}[\mathrm{y}|\mathbf{x}=\boldsymbol{x}])^2\,|\mathbf{x}=\boldsymbol{x}\right].$$

条件方差作为随机变量 $\mathbf{x}$ 的函数可记为 $\sigma^2(\mathbf{x})$。 ♣

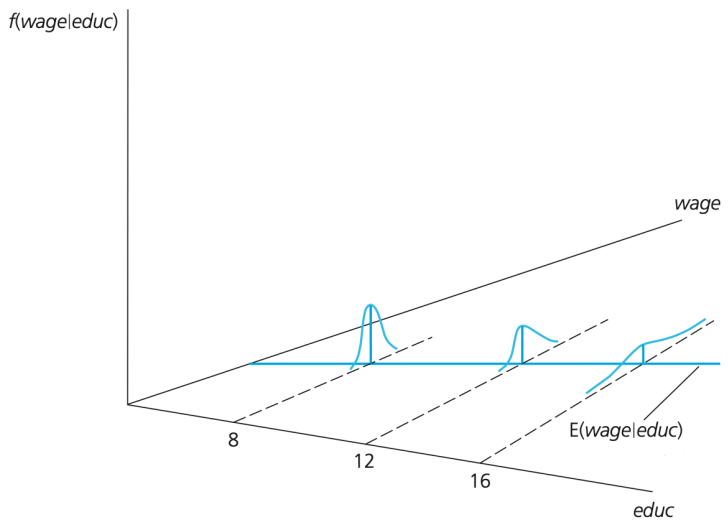

图 1.1: 条件期望与条件方差（图片来源：Wooldridge (2019) 中图 2.9）

由条件方差的定义可知条件方差也是随机变量 $\mathbf{x}$ 的函数。给定 $\mathbf{x}$ 的具体实现值 $\boldsymbol{x}$，条件期望 $\mathbb{E}[\mathrm{y}|\mathbf{x}=\boldsymbol{x}]$ 计算的是 y 的条件均值，而条件方差 $\mathrm{Var}[\mathrm{y}|\mathbf{x}=\boldsymbol{x}]$ 度量的是 y 的条件分布与其条件均值之间的偏离程度。对于对称的条件分布，条件期望给出了条件分布

的中心位置，条件方差刻画的是条件分布偏离中心位置的程度。在图 1.1 中，$\mathbf{x}$ 表示受教育年限 $educ$，y 表示工资收入 $wage$，图中分别给出了受教育年限等于 8、12 和 16 年的条件下，工资收入条件分布的均值位置（三条虚线和条件均值曲线 $\mathbb{E}[wage|educ]$ 的交点）以及条件分布的密度函数。我们发现，受教育年限越高，工资收入偏离平均工资收入的程度越大，这种现象在计量经济学中被称为条件异方差（Conditional Heteroskedasticity）。

下面我们给出一些关于条件方差的重要结论：

定理 1.4. 方差分解（Decomposition of Variance, DV）

若 $\mathbb{E}[\mathrm{y}^2] < \infty$，则有：

$$\mathrm{Var}[\mathrm{y}] = \mathbb{E}\left[\mathrm{Var}\left[\mathrm{y}|\mathbf{x}\right]\right] + \mathrm{Var}\left[\mathbb{E}\left[\mathrm{y}|\mathbf{x}\right]\right].$$

♡

证明 根据 CEF 分解性质，我们有：

$$\begin{aligned}
\mathrm{Var}[\mathrm{y}] &= \mathrm{Var}\left\{\mathbb{E}\left[\mathrm{y}|\mathbf{x}\right] + \varepsilon\right\} \\
&= \mathrm{Var}[\mathbb{E}\left[\mathrm{y}|\mathbf{x}\right]] + \mathrm{Var}(\varepsilon) && (\varepsilon \text{ 和 } \mathbb{E}[\mathrm{y}|\mathbf{x}] \text{ 不相关}) \\
&= \mathrm{Var}[\mathbb{E}\left[\mathrm{y}|\mathbf{x}\right]] + \mathbb{E}(\varepsilon^2) && (\mathbb{E}(\varepsilon) = 0) \\
&= \mathrm{Var}[\mathbb{E}\left[\mathrm{y}|\mathbf{x}\right]] + \mathbb{E}[\mathbb{E}[\varepsilon^2|\mathbf{x}]] && (\text{由 SLIE}) \\
&= \mathrm{Var}[\mathbb{E}\left[\mathrm{y}|\mathbf{x}\right]] + \mathbb{E}\left[\mathbb{E}\left[\left[\mathrm{y} - \mathbb{E}\left[\mathrm{y}|\mathbf{x}\right]\right]^2|\mathbf{x}\right]\right] && (\varepsilon = \mathrm{y} - \mathbb{E}[\mathrm{y}|\mathbf{x}]) \\
&= \mathrm{Var}[\mathbb{E}\left[\mathrm{y}|\mathbf{x}\right]] + \mathbb{E}\left[\mathrm{Var}\left[\mathrm{y}|\mathbf{x}\right]\right]. && (\text{根据条件方差的定义})
\end{aligned}$$

□

方差分解定理表明 y 的方差可以分解成 $\mathbb{E}[\mathrm{y}|\mathbf{x}]$ 的方差与 ε 的方差之和，前者是可由 $\mathbf{x}$ 解释的部分，后者是与 $\mathbf{x}$ 无关的部分。方差分解定理可以推广到 y 的条件方差，我们有如下结论：

定理 1.5. 条件方差分解（Decomposition of Conditional Variance, DCV）

若 $\mathbb{E}[\mathrm{y}^2] < \infty$，则有：

$$\mathrm{Var}(\mathrm{y}|\mathbf{x}) = \mathbb{E}[\mathrm{Var}(\mathrm{y}|\mathbf{x},\mathbf{z})|\mathbf{x}] + \mathrm{Var}[\mathbb{E}(\mathrm{y}|\mathbf{x},\mathbf{z})|\mathbf{x}].$$

♡

证明 首先，我们对 $\mathrm{Var}(\mathrm{y}|\mathbf{x})$ 进行分解：

$$\begin{aligned}
\mathrm{Var}(\mathrm{y}|\mathbf{x}) &\equiv \mathbb{E}\left\{[\mathrm{y} - \mathbb{E}(\mathrm{y}|\mathbf{x})]^2|\mathbf{x}\right\} \\
&= \mathbb{E}\left\{[\mathrm{y} - \mathbb{E}(\mathrm{y}|\mathbf{x},\mathbf{z}) + \mathbb{E}(\mathrm{y}|\mathbf{x},\mathbf{z}) - \mathbb{E}(\mathrm{y}|\mathbf{x})]^2|\mathbf{x}\right\} \\
&= \mathbb{E}\left\{[\mathrm{y} - \mathbb{E}(\mathrm{y}|\mathbf{x},\mathbf{z})]^2|\mathbf{x}\right\} + \mathbb{E}\left\{[\mathbb{E}(\mathrm{y}|\mathbf{x},\mathbf{z}) - \mathbb{E}(\mathrm{y}|\mathbf{x})]^2|\mathbf{x}\right\} \\
&\quad + 2\underbrace{\mathbb{E}\left\{[\mathrm{y} - \mathbb{E}(\mathrm{y}|\mathbf{x},\mathbf{z})] \cdot [\mathbb{E}(\mathrm{y}|\mathbf{x},\mathbf{z}) - \mathbb{E}(\mathrm{y}|\mathbf{x})]|\mathbf{x}\right\}}_{=0}.
\end{aligned}$$

由 LIE 和条件方差的定义，该分解中的第一项可表示为：

$$\mathbb{E}\left\{[\mathrm{y} - \mathbb{E}(\mathrm{y}|\mathbf{x},\mathbf{z})]^2|\mathbf{x}\right\} = \mathbb{E}\left\{\mathbb{E}\left\{[\mathrm{y} - \mathbb{E}(\mathrm{y}|\mathbf{x},\mathbf{z})]^2|\mathbf{x},\mathbf{z}\right\}|\mathbf{x}\right\} = \mathbb{E}[\mathrm{Var}(\mathrm{y}|\mathbf{x},\mathbf{z})|\mathbf{x}].$$

同理，第二项可表示为：

$$\mathbb{E}\left\{[\mathbb{E}(\mathrm{y}|\mathbf{x},\mathbf{z}) - \mathbb{E}(\mathrm{y}|\mathbf{x})]^2|\mathbf{x}\right\} = \mathbb{E}\left\{[\mathbb{E}(\mathrm{y}|\mathbf{x},\mathbf{z}) - \mathbb{E}(\mathbb{E}(\mathrm{y}|\mathbf{x},\mathbf{z})|\mathbf{x})]^2|\mathbf{x}\right\} = \mathrm{Var}[\mathbb{E}(\mathrm{y}|\mathbf{x},\mathbf{z})|\mathbf{x}].$$

第三项为 0 的推导留作习题，见习题 1.。 □

我们已经知道 CEF 是关于 y 的最小化均方误差的预测，若预测时我们使用更多的信息 $(\mathbf{x}, \mathbf{z})$，相比于 $\mathbb{E}[\mathrm{y}|\mathbf{x}]$，$\mathbb{E}[\mathrm{y}|\mathbf{x}, \mathbf{z}]$ 是否是关于 y 的更好的预测呢？由定理 1.5，我们容易得到：

$$\mathbb{E}[\mathrm{Var}(\mathrm{y}|\mathbf{x})] \geqslant \mathbb{E}[\mathrm{Var}(\mathrm{y}|\mathbf{x}, \mathbf{z})],$$

其中，

$$\mathbb{E}[\mathrm{Var}(\mathrm{y}|\mathbf{x})] = \mathbb{E}\left[\mathbb{E}\left[(\mathrm{y} - \mathbb{E}[\mathrm{y}|\mathbf{x}])^2 |\mathbf{x}\right]\right] = \mathbb{E}\left[(\mathrm{y} - \mathbb{E}[\mathrm{y}|\mathbf{x}])^2\right] = \mathrm{MSE}(\mathrm{y}; \mathbb{E}[\mathrm{y}|\mathbf{x}]),$$

同理，

$$\mathbb{E}[\mathrm{Var}(\mathrm{y}|\mathbf{x}, \mathbf{z})] = \mathrm{MSE}(\mathrm{y}; \mathbb{E}[\mathrm{y}|\mathbf{x}, \mathbf{z}]).$$

于是，我们有：

$$\mathrm{MSE}(\mathrm{y}; \mathbb{E}[\mathrm{y}|\mathbf{x}]) \geqslant \mathrm{MSE}(\mathrm{y}; \mathbb{E}[\mathrm{y}|\mathbf{x}, \mathbf{z}]).$$

从上述分析我们得到 $\mathbb{E}[\mathrm{y}|\mathbf{x}, \mathbf{z}]$ 的 MSE 不大于 $\mathbb{E}[\mathrm{y}|\mathbf{x}]$ 的 MSE，也就是说，从总体（population）的角度讲，基于更多的信息对 y 进行预测有助于减少预测的均方误差。除此之外，上述分析与如下定理的结论等价：

定理 1.6. 方差缩减定理（Variance Reduction Theorem, VRT）

若 $\mathbb{E}[\mathrm{y}^2] < \infty$，则有：

$$\mathrm{Var}[\mathrm{y}] \geqslant \mathrm{Var}[\mathrm{y} - \mathbb{E}[\mathrm{y}|\mathbf{x}]] \geqslant \mathrm{Var}[\mathrm{y} - \mathbb{E}[\mathrm{y}|\mathbf{x}, \mathbf{z}]].$$

♡

证明 记 $\varepsilon = \mathrm{y} - \mathbb{E}[\mathrm{y}|\mathbf{x}]$，则有：

$$\mathrm{Var}[\mathrm{y} - \mathbb{E}[\mathrm{y}|\mathbf{x}]] = \mathrm{Var}(\varepsilon) = \mathbb{E}(\varepsilon^2) = \mathrm{MSE}(\mathrm{y}; \mathbb{E}[\mathrm{y}|\mathbf{x}]).$$

同理，

$$\mathrm{Var}[\mathrm{y} - \mathbb{E}[\mathrm{y}|\mathbf{x}, \mathbf{z}]] = \mathrm{MSE}(\mathrm{y}; \mathbb{E}[\mathrm{y}|\mathbf{x}, \mathbf{z}]).$$

因此，第二个“$\geqslant$”得证。由定理 1.4 的证明易知 $\mathrm{Var}(\mathrm{y}) \geqslant \mathrm{Var}(\varepsilon)$，第一个“$\geqslant$”得证。□

此定理的结论实际上与我们在下一章定义的拟合优度和 R^2 的概念相对应，不同的是我们这里讨论的是总体的性质，下一章讨论的是样本的性质。

1.3 线性投影

虽然条件期望 $\mu(\mathbf{x})$ 提供了 y 的最优预测，但是 $\mu(\mathbf{x})$ 的函数形式通常是未知的。在实证分析时，一般假设 CEF 具有线性形式，但是这种假设往往不切实际。更准确的做法是将线性形式的函数设定看作是关于 CEF 的近似。为此，我们研究 CEF 的最优线性近似，即，具有最小均方误差的线性近似。

定义 1.3. 线性投影

给定 $\mathbf{x}$，y 的最优线性预测，也称为线性投影，被定义为：

$$\mathscr{P}(\mathrm{y}|\mathbf{x}) = \mathbf{x}'\boldsymbol{\beta},$$

其中，$\boldsymbol{\beta}$ 被称为线性投影系数，它最小化如下均方预测误差：

$$S(\boldsymbol{b}) = \mathbb{E}\left((\mathrm{y} - \mathbf{x}'\boldsymbol{b})^2\right).$$

即，

$$\boldsymbol{\beta} = \arg\min_{\boldsymbol{b}\in\mathbb{R}^K} S(\boldsymbol{b}).$$

♣

为了推导 $\boldsymbol{\beta}$ 的解析表达式，需要施加如下假设：

假设 1.1.

(a) $\mathbb{E}[\mathrm{y}^2] < \infty$；

(b) $\mathbb{E}\|\mathbf{x}\|^2 < \infty$；

(c) $\boldsymbol{Q}_{\mathbf{xx}} = \mathbb{E}[\mathbf{x}\mathbf{x}']$ 是正定矩阵。

定理 1.7. 线性投影的性质

(1) 线性投影系数 $\boldsymbol{\beta}$ 存在且唯一，其表达式为：

$$\boldsymbol{\beta} = (\mathbb{E}[\mathbf{x}\mathbf{x}'])^{-1}\mathbb{E}[\mathbf{x}\mathrm{y}].$$

(2) $\mathscr{P}(\mathrm{y}|\mathbf{x}) = \mathbf{x}'(\mathbb{E}[\mathbf{x}\mathbf{x}'])^{-1}\mathbb{E}[\mathbf{x}\mathrm{y}]$.

(3) 投影误差 $e = \mathrm{y} - \mathbf{x}'\boldsymbol{\beta}$ 存在，并且满足 $\mathbb{E}[e^2] < \infty$ 和 $\mathbb{E}[\mathbf{x}e] = \mathbf{0}$。特别地，若 $\mathbf{x}$ 中含有常数项，则 $\mathbb{E}(e) = 0$。

♡

证明 首先，将 $S(\boldsymbol{b})$ 进行分解：

$$S(\boldsymbol{b}) = \mathbb{E}[\mathrm{y}^2] + \boldsymbol{b}'\mathbb{E}[\mathbf{x}\mathbf{x}']\boldsymbol{b} - 2\boldsymbol{b}'\mathbb{E}[\mathbf{x}\mathrm{y}].$$

对上式两端取偏导，由最优化的一阶条件，我们有：

$$\mathbf{0} = \frac{\partial S(\boldsymbol{b})}{\partial \boldsymbol{b}}\Big|_{\boldsymbol{b}=\boldsymbol{\beta}} = -2\mathbb{E}[\mathbf{x}\mathrm{y}] + 2\mathbb{E}[\mathbf{x}\mathbf{x}']\boldsymbol{\beta}.$$

移项整理得：

$$\mathbb{E}[\mathbf{x}\mathbf{x}']\boldsymbol{\beta} = \mathbb{E}[\mathbf{x}\mathrm{y}]. \tag{1.4}$$

由期望不等式 (B.17)、Frobenius 范数定义 (A.5) 和假设 1.1 (b)，我们有：

$$\|\mathbb{E}[\mathbf{x}\mathbf{x}']\| \leqslant \mathbb{E}\|\mathbf{x}\mathbf{x}'\| = \mathbb{E}\|\mathbf{x}\|^2 < \infty.$$

同理，由期望不等式 (B.16)、柯西-施瓦茨不等式 (A.1) 和假设 1.1，我们有：

$$\|\mathbb{E}[\mathbf{x}\mathrm{y}]\| \leqslant \mathbb{E}\|\mathbf{x}\mathrm{y}\| \leqslant \left(\mathbb{E}\|\mathbf{x}\|^2\right)^{\frac{1}{2}}\left(\mathbb{E}[\mathrm{y}^2]\right)^{\frac{1}{2}} < \infty.$$

因此，$\mathbb{E}[\mathbf{x}\mathbf{x}']$ 和 $\mathbb{E}[\mathbf{x}\mathrm{y}]$ 是良定义的（well defined）。再由假设 1.1 (c) 知，$\mathbb{E}[\mathbf{x}\mathbf{x}']$ 可逆，则有：

$$\boldsymbol{\beta} = (\mathbb{E}[\mathbf{x}\mathbf{x}'])^{-1}\mathbb{E}[\mathbf{x}\mathrm{y}]. \tag{1.5}$$

由定义1.3，(2) 显然成立。

最后，我们证明 (3)：

$$
\begin{aligned}
\mathbb{E}(e^2) &= \mathbb{E}\left[\left(\mathrm{y}-\mathbf{x}'\boldsymbol{\beta}\right)^2\right] \\
&= \mathbb{E}[\mathrm{y}^2] - 2\mathbb{E}[\mathrm{y}\mathbf{x}']\boldsymbol{\beta} + \boldsymbol{\beta}'\mathbb{E}[\mathbf{x}\mathbf{x}']\boldsymbol{\beta} \\
&= \mathbb{E}[\mathrm{y}^2] - \mathbb{E}[\mathrm{y}\mathbf{x}'](\mathbb{E}[\mathbf{x}\mathbf{x}'])^{-1}\mathbb{E}[\mathbf{x}\mathrm{y}] && \text{(将式 (1.5) 的结论代入)} \\
&\leqslant \mathbb{E}[\mathrm{y}^2] < \infty. && \text{($(\mathbb{E}[\mathbf{x}\mathbf{x}'])^{-1}$ 半正定)}
\end{aligned}
$$

由式 (1.4)，我们有：

$$
\mathbf{0} = \mathbb{E}[\mathbf{x}\mathrm{y}] - \mathbb{E}[\mathbf{x}\mathbf{x}']\boldsymbol{\beta} = \mathbb{E}\left[\mathbf{x}\left(\mathrm{y}-\mathbf{x}'\boldsymbol{\beta}\right)\right] = \mathbb{E}\left[\mathbf{x}e\right].
$$

□

定理 1.7 给出了线性投影系数 $\boldsymbol{\beta}$ 的一般表达式，接下来我们考虑两种特殊情况：

- $\mathbf{x} = (1, \mathrm{x}_1)'$，在此种情况下，我们有：

$$
\mathrm{y} = \alpha + \beta_1 \mathrm{x}_1 + e,
$$

其中，α 和 β_1 分别是常数项 1 和 x_1 对应的线性投影系数。容易证明：

$$
\beta_1 = \frac{\mathrm{Cov}(\mathrm{y}, \mathrm{x}_1)}{\mathrm{Var}(\mathrm{x}_1)}, \quad \alpha = \mathbb{E}(\mathrm{y}) - \beta_1 \mathbb{E}(\mathrm{x}_1). \tag{1.6}
$$

- $\mathbf{x} = (1, \mathrm{x}_1, \mathrm{x}_2, \ldots, \mathrm{x}_K)'$，$K > 1$，在此种情况下，我们有：

$$
\beta_k = \frac{\mathrm{Cov}(\mathrm{y}, \tilde{\mathrm{x}}_k)}{\mathrm{Var}(\tilde{\mathrm{x}}_k)}, \tag{1.7}
$$

其中，β_k 是 x_k 对应的线性投影系数，$\tilde{\mathrm{x}}_k$ 是 x_k 对所有其他协变量投影得到的投影误差。下面我们对式 (1.7) 进行证明。

证明 首先，将 y 写成线性投影加投影误差的形式：

$$
\mathrm{y} = \alpha + \beta_1 \mathrm{x}_1 + \cdots + \beta_k \mathrm{x}_k + \cdots + \beta_K \mathrm{x}_K + e.
$$

将上式代入 $\mathrm{Cov}(\mathrm{y}, \tilde{\mathrm{x}}_k)$，注意到，由习题 4.，我们有：

- $\mathrm{Cov}(\tilde{\mathrm{x}}_k, e) = 0$，因为 $\tilde{\mathrm{x}}_k$ 是 $\mathbf{x}$ 各个分量的线性组合，e 与 $\mathbf{x}$ 不相关；
- $\mathrm{Cov}(\tilde{\mathrm{x}}_k, \mathrm{x}_j) = 0$，$j \neq k$，因为 $\tilde{\mathrm{x}}_k$ 是线性投影误差，与 x_j 不相关；
- $\mathrm{Cov}(\tilde{\mathrm{x}}_k, \mathrm{x}_k) = \mathrm{Cov}(\tilde{\mathrm{x}}_k, \tilde{\mathrm{x}}_k + \mathbf{x}_{-k}'\boldsymbol{\beta}_{-k}) = \mathrm{Var}(\tilde{\mathrm{x}}_k)$，其中，$\mathbf{x}_{-k}$ 表示除 x_k 之外的所有协变量构成的向量，即，$\mathbf{x}_{-k} = (1, \mathrm{x}_1, \ldots, \mathrm{x}_{k-1}, \mathrm{x}_{k+1}, \ldots, \mathrm{x}_K)'$，$\boldsymbol{\beta}_{-k}$ 是 x_k 对 $\mathbf{x}_{-k}$ 进行投影得到的线性投影系数。

于是，

$$
\frac{\mathrm{Cov}(\mathrm{y}, \tilde{\mathrm{x}}_k)}{\mathrm{Var}(\tilde{\mathrm{x}}_k)} = \frac{\beta_k \mathrm{Var}(\tilde{\mathrm{x}}_k)}{\mathrm{Var}(\tilde{\mathrm{x}}_k)} = \beta_k.
$$

□

我们把第一种情况称为一元线性投影或简单线性投影，把第二种情况称为多元线性投影。由定理 1.7 中 (3) 的结论，线性投影误差 $\tilde{\mathrm{x}}_k$ 与 $\mathbf{x}_{-k}$ 正交，即，$\mathbb{E}[\mathbf{x}_{-k}\tilde{\mathrm{x}}_k] = \mathbf{0}$。式 (1.7) 说明多元线性投影中 x_k 的线性投影系数可以通过如下方式得到：首先，将 x_k 对 $\mathbf{x}_{-k}$ 做投影并保留误差 $\tilde{\mathrm{x}}_k$，目的是从 x_k 中剔除所有其他协变量的影响；然后，将 y 对误差 $\tilde{\mathrm{x}}_k$ 和常数项 1 做简单线性投影得到系数 β_k。

计量分析（特别是应用于公司金融等实证方向的微观计量分析）的主要目的之一是

推断变量之间的因果关系。建立因果关系的核心是“控制其他所有相关因素不变”，拉丁语为 ceteris paribus。我们将在下一章多元线性回归中看到“控制其他所有相关因素不变”和“剔除所有其他因素的影响”是一致的。式 (1.7) 是从总体的角度讨论线性投影系数，下一章将从样本的角度讨论线性投影系数的估计。式 (1.7) 对应着多元线性回归的 Frisch-Waugh 定理。

关于线性投影，我们还有如下重要的结论：

定理 1.8. 线性投影与 CEF

(1) 假设 CEF 是 $\mathbf{x}$ 的线性函数，则 $\mu(\mathbf{x}) = \mathscr{P}(\mathrm{y}|\mathbf{x})$.

(2) 给定 $\mathbf{x}$，$\mathscr{P}(\mathrm{y}|\mathbf{x})$ 是关于 y 的最小化 MSE 的最优线性预测。

(3) $\mathscr{P}(\mathrm{y}|\mathbf{x})$ 是 $\mathbb{E}[\mathrm{y}|\mathbf{x}]$ 的最优（最小化 MSE）线性近似，即，

$$\boldsymbol{\beta} = \arg\min_{\boldsymbol{b}} \mathbb{E}\left\{\left(\mathbb{E}\left[\mathrm{y}|\mathbf{x}\right] - \mathbf{x}'\boldsymbol{b}\right)^2\right\}.$$

♡

证明 首先，证明 (1)，假设 $\mu(\mathbf{x}) = \mathbf{x}'\boldsymbol{\beta}^*$，由 CEF 分解性质，$\mathbb{E}[\mathbf{x}(\mathrm{y} - \mathbb{E}[\mathrm{y}|\mathbf{x}])] = \mathbf{0}$，将 $\mu(\mathbf{x}) = \mathbf{x}'\boldsymbol{\beta}^*$ 代入并将 $\boldsymbol{\beta}^*$ 解出，得到 $\boldsymbol{\beta}^* = \boldsymbol{\beta}$。

由定义1.3知 (2) 显然成立。最后，我们证明 (3)，注意到：

$$\begin{aligned}(\mathrm{y} - \mathbf{x}'\boldsymbol{b})^2 &= \left\{(\mathrm{y} - \mathbb{E}[\mathrm{y}|\mathbf{x}]) + (\mathbb{E}[\mathrm{y}|\mathbf{x}] - \mathbf{x}'\boldsymbol{b})\right\}^2 \\ &= (\mathrm{y} - \mathbb{E}\left[\mathrm{y}|\mathbf{x}\right])^2 + \left(\mathbb{E}[\mathrm{y}|\mathbf{x}] - \mathbf{x}'\boldsymbol{b}\right)^2 + 2\left(\mathrm{y} - \mathbb{E}[\mathrm{y}|\mathbf{x}]\right)\cdot\left(\mathbb{E}[\mathrm{y}|\mathbf{x}] - \mathbf{x}'\boldsymbol{b}\right).\end{aligned}$$

所以，最小化 $\mathbb{E}\left\{\left(\mathbb{E}\left[\mathrm{y}|\mathbf{x}\right] - \mathbf{x}'\boldsymbol{b}\right)^2\right\}$ 等价于最小化 $\mathbb{E}\left\{(\mathrm{y} - \mathbf{x}'\boldsymbol{b})^2\right\}$，由 (2) 知，(3) 成立。□

我们在前文中多次提到 CEF 提供了关于 y 的最优预测，而定理 1.8 的结论 (2) 说明线性投影提供了关于 y 的最优线性预测。除此之外，由于 CEF 的形式是未知的，我们一般假设其具有线性的形式，从定理 1.8 的结论 (3) 中可以看出这种假设具有一定的合理性。线性投影系数 $\boldsymbol{\beta}$ 也被称为总体回归系数（population regression coefficient），这是因为线性回归和线性投影具有密不可分的关系。实际上，线性回归是利用样本估计线性投影系数。

1.4 饱和回归模型

饱和回归模型是具有离散解释变量的回归模型，对解释变量取值的每一个可能组合，饱和回归模型都为该组合对应的条件期望引入一个额外的参数。我们将要看到饱和回归模型具有线性的 CEF。定理 1.8 的结论 (1) 表明：如果 CEF 具有线性形式，则 CEF 等价于线性投影。因此，对于饱和回归模型，我们有：$\mu(\mathbf{x}) = \mathscr{P}(\mathrm{y}|\mathbf{x})$。

首先，我们考虑只含有一个二值（binary）解释变量的情况，二值变量又被称为虚拟（dummy）变量。假设二值变量 x 代表性别，CEF 只能有两个不同的取值：

$$\mathbb{E}[\,\mathrm{y}\mid \text{性别}\,] = \begin{cases} \mu_0 & \text{如果} \quad \text{性别} = \text{男性}, \\ \mu_1 & \text{如果} \quad \text{性别} = \text{女性}. \end{cases}$$

为了便于数学上的处理，我们引入虚拟变量 x_1，满足：

$$\mathrm{x}_1 = \begin{cases} 0 & \text{如果} \quad \text{性别} = \text{男性}, \\ 1 & \text{如果} \quad \text{性别} = \text{女性}. \end{cases}$$

于是，$\mathbb{E}[\mathrm{y}|\mathrm{x}_1] = \beta_0 + \beta_1 \mathrm{x}_1$，其中，$\beta_0 = \mu_0$，$\beta_1 = \mu_1 - \mu_0$。此时，饱和回归模型可以表示为：

$$\mathrm{y} = \beta_0 + \beta_1 \mathrm{x}_1 + \varepsilon.$$

我们通常称 β_0 为截距，它表示男性总体 y 的期望；称 β_1 为斜率，它表示女性总体和男性总体 y 的期望之差。

类似地，若解释变量是多值离散变量 $\mathrm{x}_1 \in \{0, 1, 2, \ldots, \tau\}$，关于 x_1 的饱和回归模型可以表示为：

$$\mathrm{y} = \beta_0 + \beta_1 \mathrm{d}_{11} + \beta_2 \mathrm{d}_{12} + \cdots + \beta_\tau \mathrm{d}_{1\tau} + \varepsilon,$$

其中，$\mathrm{d}_{1j} = \mathbf{1}[\mathrm{x}_1 = j]$，$j = 1, 2, \ldots, \tau$，$\mathbf{1}[A]$ 是示性函数，条件 A 成立时取值为 1，否则取值为 0。由上式容易验证：

$$\beta_0 = \mathbb{E}[\mathrm{y}|\mathrm{x}_1 = 0], \quad \beta_j = \mathbb{E}[\mathrm{y}|\mathrm{x}_1 = j] - \mathbb{E}[\mathrm{y}|\mathrm{x}_1 = 0].$$

接下来，我们考虑含有两个虚拟解释变量的情况。假设虚拟变量 x_2 满足：

$$\mathrm{x}_2 = \begin{cases} 1 & \text{如果} \quad \text{具有大学学历}, \\ 0 & \text{如果} \quad \text{没有大学学历}. \end{cases}$$

以 x_1 和 x_2 为条件的条件期望函数有四种可能的取值：

$$\mathbb{E}[\mathrm{y}|\mathrm{x}_1, \mathrm{x}_2] = \begin{cases} \mu_{00} & \text{如果} \quad \mathrm{x}_1 = 0, \mathrm{x}_2 = 0 \\ \mu_{01} & \text{如果} \quad \mathrm{x}_1 = 0, \mathrm{x}_2 = 1 \\ \mu_{10} & \text{如果} \quad \mathrm{x}_1 = 1, \mathrm{x}_2 = 0 \\ \mu_{11} & \text{如果} \quad \mathrm{x}_1 = 1, \mathrm{x}_2 = 1. \end{cases}$$

此条件期望函数可以写成 $(\mathrm{x}_1, \mathrm{x}_2, \mathrm{x}_1\mathrm{x}_2)'$ 的线性函数：

$$\mathbb{E}[\mathrm{y}|\mathrm{x}_1, \mathrm{x}_2] = \beta_0 + \beta_1 \mathrm{x}_1 + \beta_2 \mathrm{x}_2 + \beta_3 \mathrm{x}_1 \mathrm{x}_2,$$

其中，$\beta_0 = \mu_{00}$，$\beta_1 = \mu_{10} - \mu_{00}$，$\beta_2 = \mu_{01} - \mu_{00}$，$\beta_3 = \mu_{11} - \mu_{10} - \mu_{01} + \mu_{00}$。相应的饱和回归模型可以表示为：

$$\mathrm{y} = \beta_0 + \beta_1 \mathrm{x}_1 + \beta_2 \mathrm{x}_2 + \beta_3 \mathrm{x}_1 \mathrm{x}_2 + \varepsilon.$$

在上式中，x_1 和 x_2 被称为主项，$\mathrm{x}_1\mathrm{x}_2$ 被称为交互项。假设变量 y 表示工资，则 β_0 表示没有大学学历的男性的期望工资；β_1 衡量没有大学学历的女性与男性的期望工资的差异；β_2 衡量大学学历对男性期望工资的影响；$\beta_3 = (\mu_{11} - \mu_{10}) - (\mu_{01} - \mu_{00})$ 衡量的是大学学历对女性期望工资的影响与大学学历对男性期望工资的影响之间的差异。

最后，我们考虑二值虚拟变量和多值离散变量构成的饱和回归模型。假设 x_1 依然是表示性别的虚拟变量；x_2 表示受教育年限，值域为 $\{0, 1, 2, \ldots, \tau\}$。相对于 x_1 和 x_2 的饱

和模型可以表示为：

$$\mathrm{y} = \beta_0 + \sum_{j=1}^{\tau} \beta_j \mathrm{d}_{2j} + \gamma \mathrm{x}_1 + \sum_{j=1}^{\tau} \delta_j (\mathrm{d}_{2j} \mathrm{x}_1) + \varepsilon,$$

其中，$\mathrm{d}_{2j} = \mathbf{1}[\mathrm{x}_2 = j]$，$j = 1, 2, \ldots, \tau$。通常，在回归分析中，若线性模型中含有交互项，则相应的主项也应该出现在回归模型中。考虑如下去掉受教育年限主项的回归模型：

$$\mathrm{y} = \beta_0 + \gamma \mathrm{x}_1 + \sum_{j=1}^{\tau} \delta_j (\mathrm{d}_{2j} \mathrm{x}_1) + \varepsilon,$$

若 y 表示工资，这一模型只允许女性通过改变受教育年限影响工资收入，男性的工资收入和受教育年限无关，这与实际情况不符。

1.5 总结

本章详细介绍了条件期望、条件方差和线性投影的定义和主要性质，为后续计量经济学知识的学习奠定理论基础。通过本章的学习读者应该理解为什么条件期望是计量经济学研究的核心对象之一，这涉及到 CEF 的分解性质和预测性质。此外，迭代期望律在计量方法理论性质的推导中经常被使用，本章给出了（简单）迭代期望律的详细证明，并在其他定理和性质的证明过程中反复使用此定律，希望读者能够掌握此定律并熟练使用。

本章的另外一个重要内容是线性投影。线性投影是条件期望函数的最优线性近似，对变量 $\mathbf{x}$ 做投影得到的误差 e 与 $\mathbf{x}$ 正交，即，$\mathbb{E}[\mathbf{x}e] = \mathbf{0}$。我们在下一章将会看到最小二乘估计量满足的正则方程（normal equation）实际上是此正交条件的样本形式。因此，最小二乘估计量实际上估计的是线性投影系数，对线性投影系数的理解有助于读者深入理解线性回归。这里读者应该注意区分 CEF 的分解性质中出现的 ε 和线性投影误差 e，前者满足 $\mathbb{E}[\varepsilon|\mathbf{x}] = 0$；但是，通常情况下，$\mathbb{E}[e|\mathbf{x}] = 0$ 并不成立，除非 CEF 为线性形式。

本章的写作借鉴了如下优秀计量教材的相关章节：Angrist 和 Pischke 两位教授合著的 *Mostly Harmless Econometrics: An Empiricist's Companion* 的第三章；Hansen 教授的线上教材手稿 *Econometrics* 的第二章；Wooldridge 教授的 *Econometric Analysis of Cross Section and Panel Data* 的第二章。推荐想进一步深入学习本章内容的读者阅读。

这里借用 *Mostly Harmless Econometrics: An Empiricist's Companion* 第三章的标题 "Making Regression Make Sense" 来说明本章的目的：我们通过对总体性质的讨论，厘清回归分析感兴趣的参数，在利用具体的样本对这些参数进行估计之前，先明确在总体中这些参数的具体含义。

第一章习题

1. 证明：定理 1.5 的证明过程中关于 $\mathrm{Var}(\mathrm{y}|\mathbf{x})$ 的分解表达式的第三项为 0。
2. 试用 CEF 的分解性质证明定理 1.6。
3. 证明：在假设 1.1 成立的条件下，定理 1.7 中定义的投影误差 e 满足 $\mathbb{E}[\mathbf{x}e] \prec \infty$。[2]
4. 证明：在定理 1.7 中，若 $\mathbf{x}$ 中含有常数项，则 $\mathbf{x}$ 与 e 不相关。

[2] 不能利用 $\mathbb{E}[\mathbf{x}e] = \mathbf{0}$ 的结论。

5. 证明：利用定理 1.7 中 (1) 的结论证明式 (1.6) 成立。

第二章 多元线性回归

内容提要

- 古典线性回归模型 2.1
- 普通最小二乘法 2.2
- Frisch-Waugh 定理 2.1
- 非中心化 R^2 2.3
- 中心化 R^2 2.4
- 无偏性 2.2
- Gauss-Markov 定理 2.2
- t 统计量 2.5
- F 统计量 2.6
- 多重共线性 2.4
- 广义最小二乘 2.8
- 一致性 2.6
- 渐近正态性 2.6
- 稳健 t 统计量 2.6

多元线性回归是计量经济学中最常用的计量模型之一，本章第一节将介绍古典线性回归模型（Classical Linear Regression Model, CLRM）的假设条件以及估计此模型的普通最小二乘法（Ordinary Least Squares, OLS）；第二节讨论线性回归模型的拟合优度并推导古典假设下最小二乘估计量的有限样本性质（或称为小样本性质）；第三节介绍古典线性回归模型的假设检验问题；第四节讨论多重共线性问题；在第五节中，我们放松球形扰动假设并讨论广义最小二乘；在第六节中，我们建立最小二乘估计量的大样本性质和大样本统计推断理论。想深入了解本章内容可以进一步参考 Hansen (2022a) 和 Hayashi (2000)。

计量实证分析常见的数据结构有：横截面（cross-sectional）数据、时间序列（time series）数据、混合（pooled）横截面数据和面板（panel）数据。横截面数据是指在给定的时间点上由不同个体的各个变量的样本观测值构成的数据，例如，2022 年上半年我国各个省份的人均 GDP、平均工资水平和就业率构成的数据集便是横截面数据。这里需要注意的是，即使收集各省数据时存在时间上的差别（如，2022 年上半年的不同星期或月份），我们仍然可以将其看作是横截面数据。时间序列数据是指由同一个体的各个变量在时间维度上的观测值构成的数据，例如，贵州茅台 2022 年上半年的日开盘价、日收盘价和日成交量构成的数据集是时间序列数据。在记录横截面数据时，不同个体的记录顺序是无关紧要的，但是，时间序列数据要按照时间的先后顺序进行记录。将不同时间点的相同变量的横截面数据按时间顺序混合在一起便得到混合横截面数据。在此类数据中，不同时间点的横截面个体可以不同。面板数据是由每一横截面个体在相同时间点上的时间序列构成的数据。面板数据不同于混合横截面数据，面板数据不同时间点的横截面个体是相同的。例如，我国各省份 2012-2021 年的人均 GDP、平均工资水平和就业率构成的数据是面板数据。

本章介绍的小样本性质适用于横截面数据。通常横截面数据假设横截面个体之间是独立同分布的，但是，我们在推导小样本性质时没有施加此假设。此外，本章介绍的大样本性质适用于横截面数据和时间序列数据。关于（静态）面板数据的详细讨论见本书第 5 章。

2.1 古典线性回归模型

本章主要讨论多元线性回归模型中的参数的估计和假设检验问题，为此，我们假设有样本量为 n 的样本 $(y_i, \mathbf{x}_i)_{i=1}^n$，其中，$\mathbf{x}_i$ 为 K 维列向量。计量经济学中所指的模型是对解释变量 $\mathbf{x}_i$ 和被解释变量 y_i 的联合分布施加的一组假设。古典线性回归模型满足的假设由下面定义给出：

定义 2.1. 古典线性回归模型（Classical Linear Regression Model, CLRM）

古典线性回归模型满足如下假设：

假设 2.1：（线性假设）

$$y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_K x_{iK} + \varepsilon_i = \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i, \quad (i = 1, 2, \ldots, n), \tag{2.1}$$

其中，$\mathbf{x}_i = (x_{i1}, \ldots, x_{iK})'$；$\boldsymbol{\beta} = (\beta_1, \ldots, \beta_K)'$ 是待估计的未知参数；ε_i 是不可观测的误差项；在含有常数项的回归模型中，$x_{i1} = 1$。若再定义：

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} \mathbf{x}_1' \\ \mathbf{x}_2' \\ \vdots \\ \mathbf{x}_n' \end{bmatrix} = \begin{bmatrix} x_{11} & \cdots & x_{1K} \\ x_{21} & \cdots & x_{2K} \\ \vdots & \cdots & \vdots \\ x_{n1} & \cdots & x_{nK} \end{bmatrix},$$

则有：

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}.$$

假设 2.2：（严格外生性）

$$\mathbb{E}(\varepsilon_i|\mathbf{X}) = 0, \quad (i = 1, 2, \ldots, n). \tag{2.2}$$

假设 2.3：（无完全共线性）

$$\text{矩阵 } \mathbf{X} \text{ 列满秩，即，} \operatorname{rank}(\mathbf{X}) = K. \tag{2.3}$$

假设 2.4：（球形扰动）

$$\mathbb{E}(\varepsilon_i^2|\mathbf{X}) = \sigma^2 > 0, \quad (i = 1, 2, \ldots, n) \tag{2.4}$$

$$\mathbb{E}(\varepsilon_i\varepsilon_j|\mathbf{X}) = 0, \quad (i, j = 1, 2, \ldots, n;\ i \neq j). \tag{2.5}$$

上述两个条件可以统一记为：

$$\mathbb{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'|\mathbf{X}) = \operatorname{Var}(\boldsymbol{\varepsilon}|\mathbf{X}) = \sigma^2\mathbf{I}_n, \tag{2.6}$$

其中，$\mathbf{I}_n$ 为 $n \times n$ 的单位阵。 ♣

关于假设的说明

在式 (2.1) 中，$\beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_K x_{iK}$ 称为总体回归函数，$\boldsymbol{\beta}$ 称为回归系数。假设 2.1 的线性假设实际上要求总体回归函数关于回归系数 $\boldsymbol{\beta}$ 是线性的。考虑如下方程：

$$y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i2}^2 + \varepsilon_i.$$

虽然此方程是变量 x_{i2} 的非线性函数，但是我们可以定义 $x_{i3} = x_{i2}^2$，将上述方程改写为：

$$y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \varepsilon_i.$$

改写后的新方程关于 $\boldsymbol{\beta}$ 是线性的，满足假设 2.1。由于 x_{i2} 是可观测的，x_{i3} 可由观测值计算得出，所以 $\boldsymbol{\beta}$ 的估计不受影响。再考虑如下方程：

$$y_i = \beta_1 x_{i1} + \frac{\beta_2}{\beta_3} x_{i2} + \varepsilon_i.$$

此方程不满足假设 2.1，无法基于样本数据估计 β_2 和 β_3。

在总体回归函数中，对于 $k \in \{1, 2, \ldots, K\}$，$x_{ik}$ 的系数 β_k 衡量的是 x_{ik} 对 y_i 的偏效应（partial effect），即，在保持其他解释变量不变的情况下，x_{ik} 变化一个单位引起的 y_i 的变化，利用微积分中偏导数的记号可以记为 $\partial y_i / \partial x_{ik} = \beta_k$。在实证研究中，我们经常会对解释变量或被解释变量取对数，此时回归系数的解释要做相应的调整。考虑如下简单线性回归：[1]

$$\log(wage)_i = \alpha + \beta_1 educ_i + \varepsilon_i,$$

其中，$\log(wage)_i$ 表示对数工资，$educ_i$ 表示受教育年限。因为此模型里只有被解释变量取了对数，所以它被称为“对数-水平（log-level）”模型。注意到：

$$\frac{d(\log(wage)_i)}{d(educ_i)} = \frac{d(wage_i)/wage_i}{d(educ_i)} = \beta_1.$$

于是，受教育年限增加一年导致的工资百分比变化为 $100 \cdot \beta_1\%$，$100 \cdot \beta_1$ 称为工资相对于受教育年限的半弹性。例如，若 $\beta_1 = 0.067$，则增加一年的教育会使工资增加 6.7%。

除“对数-水平”模型外，常见的还有“对数-对数（log-log）”模型。考虑如下模型：

$$\log(salary)_i = \alpha + \beta_1 \log(sale)_i + \varepsilon_i,$$

其中，$\log(salary)_i$ 表示公司高管薪酬对数，$\log(sale)_i$ 表示公司销售收入对数。因为

$$\frac{d(\log(salary)_i)}{d(\log(sale)_i)} = \frac{d(salary_i)/salary_i}{d(sale_i)/sale_i} = \beta_1,$$

所以，这里的 β_1 具有弹性的解释，即，销售收入增加 1%，高管薪酬变化 $\beta_1\%$。例如，若 $\beta_1 = 0.2$，则销售收入增加 1% 会使高管薪酬增加 0.2%。

假设 2.2 也可以表示为

$$\mathbb{E}(\varepsilon_i \mid \mathbf{x}_1, \ldots, \mathbf{x}_n) = 0, \quad (i = 1, 2, \ldots, n),$$

即，以所有的观测变量 $\mathbf{x}_1, \ldots, \mathbf{x}_n$ 为条件，ε_i 的条件期望为 0。由此假设出发，我们可以推导出如下重要结论：

- 误差项的无条件期望是 0：$\mathbb{E}(\varepsilon_i) = 0$，$i = 1, 2, \ldots, n$；

  证明 由 SLIE，$\mathbb{E}[\mathbb{E}(\varepsilon_i|\mathbf{X})] = \mathbb{E}(\varepsilon_i) = 0$。 □

- 解释变量与误差项正交：$\mathbb{E}(x_{jk}\varepsilon_i) = 0$，$i, j = 1, 2, \ldots, n$; $k = 1, \ldots, K$；

  证明 首先，由 LIE，假设 2.2 可以推出

$$\mathbb{E}(\varepsilon_i|x_{jk}) = \mathbb{E}[\mathbb{E}(\varepsilon_i|\mathbf{X})|x_{jk}] = 0.$$

[1] 总体回归函数中只含有常数项和一个解释变量的线性回归模型称为简单线性回归模型，也称为一元线性回归模型。

因此，我们有：

$$\begin{aligned}\mathbb{E}(x_{jk}\varepsilon_i) &= \mathbb{E}[\mathbb{E}(x_{jk}\varepsilon_i|x_{jk})] && \text{(由 SLIE)}\\ &= \mathbb{E}[x_{jk}\mathbb{E}(\varepsilon_i|x_{jk})] && \text{(由定理 1.3)}\\ &= 0.\end{aligned}$$

□

- 解释变量与误差项不相关：$\mathrm{Cov}(\varepsilon_i, x_{jk}) = 0$，$i, j = 1, 2, \ldots, n$; $k = 1, \ldots, K$。

证明

$$\begin{aligned}\mathrm{Cov}(\varepsilon_i, x_{jk}) &= \mathbb{E}(x_{jk}\varepsilon_i) - \mathbb{E}(x_{jk})\mathbb{E}(\varepsilon_i) && \text{(协方差计算公式)}\\ &= \mathbb{E}(x_{jk}\varepsilon_i) && \text{(因为 } \mathbb{E}(\varepsilon_i) = 0\text{)}\\ &= 0.\end{aligned}$$

□

从上述结论中我们可以看到严格外生性不仅要求同一观测的解释变量和误差项彼此正交（$\mathbb{E}(x_{ik}\varepsilon_i) = 0$），还要求来自不同观测的解释变量和误差项彼此正交（对所有的 $j \neq i$，$\mathbb{E}(x_{jk}\varepsilon_i) = 0$）。这一点对于时间序列模型来说是无法满足的，例如考虑如下一阶自回归模型（First-order Autoregressive Model, AR(1)）：

$$y_i = \beta_0 + \beta_1 y_{i-1} + \varepsilon_i.$$

在此模型中，我们假设解释变量 y_{i-1} 和误差项正交：$\mathbb{E}(y_{i-1}\varepsilon_i) = 0$，以及 $\mathbb{E}(\varepsilon_i) = 0$，则有：

$$\begin{aligned}\mathbb{E}(y_i\varepsilon_i) &= \mathbb{E}[(\beta_0 + \beta_1 y_{i-1} + \varepsilon_i)\varepsilon_i]\\ &= \beta_0\mathbb{E}(\varepsilon_i) + \beta_1\mathbb{E}(y_{i-1}\varepsilon_i) + \mathbb{E}(\varepsilon_i^2)\\ &= \mathbb{E}(\varepsilon_i^2).\end{aligned}$$

除非误差项为 0，否则上式将不为 0。如果我们把变量的下标理解成时间，相对于 ε_i，y_i 是下一期的解释变量。因此，解释变量与过去期的误差项不满足正交条件，严格外生性不成立。在本章第四节中我们会放松严格外生假设使其适用于时间序列模型。

实际上，只要 $\mathbb{E}(\varepsilon_i \mid \mathbf{x}_1, \ldots, \mathbf{x}_n) = c$，其中，$c$ 为某一常数，解释变量就与误差项不相关。在含有常数项的多元回归模型中，假设 2.2 要求 $c = 0$ 并不是一个过于强的假设。因为，若 $c \neq 0$，我们可以重新定义常数项和误差项，从而使假设 2.2 成立。具体地，

$$\begin{aligned}y_i &= \beta_1 + \beta_2 x_{i2} + \cdots + \beta_K x_{iK} + \varepsilon_i\\ &= (\beta_1 + c) + \beta_2 x_{i2} + \cdots + \beta_K x_{iK} + (\varepsilon_i - c)\\ &= \beta_1^* + \beta_2 x_{i2} + \cdots + \beta_K x_{iK} + \varepsilon_i^*,\end{aligned}$$

其中，$\beta_1^* = \beta_1 + c$，$\varepsilon_i^* = \varepsilon_i - c$。重新定义的误差项 ε_i^* 满足假设 2.2。

此外，在假设 2.2 成立的情况下，容易验证：

$$\mathbb{E}(y_i|\mathbf{x}_i) = \mathbf{x}_i'\boldsymbol{\beta}.$$

于是，CEF 具有线性形式，由定理 1.8，$\boldsymbol{\beta}$ 为 y_i 对 $\mathbf{x}_i$ 做投影的线性投影系数。我们已经

知道投影误差正交于 $\mathbf{x}_i$，而 $\mathbf{x}_i$ 与回归模型的误差项 ε_i 正交是由假设 2.2 推出的，严格外生性比正交条件的约束更强。

假设 2.3 要求 $\mathbf{X}$ 列满秩，即，$\mathbf{X}$ 的任何一列都不能表示成其他列的线性组合。此外，因为 $\mathbf{X}$ 是 $n \times K$ 维矩阵，假设 2.3 成立需要 $n \geqslant K$。若此假设不满足，我们称解释变量出现完全共线性问题，此时最小二乘估计量无法计算。对于简单线性回归模型来说，假设 2.3 要求解释变量的数据不是常数，否则解释变量和回归中的常数项成比例导致数据矩阵 $\mathbf{X}$ 不列满秩。

“虚拟变量陷阱”也是导致完全共线性问题的常见例子，考虑如下模型：

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i,$$

其中，

$$x_{i1} = \begin{cases} 0 & \text{如果} \quad \text{性别 = 男性}, \\ 1 & \text{如果} \quad \text{性别 = 女性}. \end{cases} \qquad x_{i2} = \begin{cases} 1 & \text{如果} \quad \text{性别 = 男性}, \\ 0 & \text{如果} \quad \text{性别 = 女性}. \end{cases}$$

因为，对任意观测 i，$x_{i1} + x_{i2} = 1$，所以回归中的常数项可以表示为 x_{i1} 和 x_{i2} 的线性组合。一般地，若某一类别变量有 m 个不同的取值，将此变量加入回归中应该引入 $m-1$ 个虚拟变量；若引入 m 个虚拟变量，常数项可以表示成这些虚拟变量的和，从而导致假设 2.3 不成立。

Stata 会自动检测完全共线性问题，若问题存在，共线性的变量中的一个会被自动剔除。然而，在实证分析中，值得讨论的并不是完全共线性问题。假设 2.3 排除了解释变量间的完全的线性关系，但是解释变量间可能存在高度的（不完全的）线性相关性，这种相关性可能会导致最小二乘估计量的方差无限放大。我们将因为解释变量高度线性相关影响回归系数估计准确性的问题称为多重共线性（multicollinearity）问题。在本章 2.4 节我们会进一步讨论多重共线性问题。

式 (2.6) 假设 $\boldsymbol{\varepsilon}$ 的条件方差和单位阵 $\mathbf{I}_n$ 成正比，因此假设 2.4 被称为球形扰动假设。由严格外生性假设，我们有：

$$\begin{aligned} \mathrm{Var}(\varepsilon_i|\mathbf{X}) &= \mathbb{E}(\varepsilon_i^2|\mathbf{X}) - \mathbb{E}(\varepsilon_i|\mathbf{X})^2 \\ &= \mathbb{E}(\varepsilon_i^2|\mathbf{X}). \end{aligned}$$

因此，式 (2.4) 说明误差项满足条件同方差（conditional homoskedasticity）。同理，

$$\mathrm{Cov}(\varepsilon_i, \varepsilon_j|\mathbf{X}) = \mathbb{E}(\varepsilon_i\varepsilon_j|\mathbf{X}) = 0, \quad (i, j = 1, 2, \ldots, n;\ i \neq j).$$

因此，式 (2.5) 说明不同观测误差之间满足条件不相关。在时间序列模型中，上式表示不同时刻的误差项之间不存在条件相关性，这种情况被称为误差项无自相关（serial correlation）。

事实上，条件同方差是一个很强的假设，在很多情况下并不满足。例如，由定理 1.8，我们知道总体回归函数是对 CEF 的最优线性近似，即使数据真的是同方差，但 CEF 是非线性的，回归模型的误差项也将是异方差的。具体地，

$$\begin{aligned} \mathbb{E}(\varepsilon_i^2 \mid \mathbf{x}_i) &= \mathbb{E}[(y_i - \mathbf{x}_i'\boldsymbol{\beta})^2 \mid \mathbf{x}_i] \\ &= \mathbb{E}\left\{\left[y_i - \mathbb{E}(y_i \mid \mathbf{x}_i) + \mathbb{E}(y_i \mid \mathbf{x}_i) - \mathbf{x}_i'\boldsymbol{\beta}\right]^2 \mid \mathbf{x}_i\right\} \end{aligned}$$

$$= \quad \mathrm{Var}(y_i \mid \mathbf{x}_i) + \left[\mathbb{E}(y_i \mid \mathbf{x}_i) - \mathbf{x}_i'\boldsymbol{\beta}\right]^2.$$

即使 $\mathrm{Var}(y_i \mid \mathbf{x}_i) = \sigma^2$，如果 CEF 是非线性的，第二项将不等于 0，而是 $\mathbf{x}_i$ 的函数。

在 CEF 是线性的情况下，误差项也可能是异方差的。例如，考虑线性概率模型：

$$y_i = \alpha + \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i, \quad \mathbb{E}(\varepsilon_i \mid \mathbf{x}_i) = 0,$$

其中，y_i 是二值的。此模型的误差项是异方差的：

$$\begin{aligned}
\mathbb{E}(\varepsilon_i^2 \mid \mathbf{x}_i) &= \mathbb{E}\left\{[y_i - \mathbb{E}(y_i \mid \mathbf{x}_i)]^2 \mid \mathbf{x}_i\right\} \\
&= \mathbb{E}\left\{[y_i - \mathbb{P}(y_i = 1 \mid \mathbf{x}_i)]^2 \mid \mathbf{x}_i\right\} \\
&= [1 - \mathbb{P}(y_i = 1 \mid \mathbf{x}_i)]^2\, \mathbb{P}(y_i = 1 \mid \mathbf{x}_i) + [\mathbb{P}(y_i = 1 \mid \mathbf{x}_i)]^2\, [1 - \mathbb{P}(y_i = 1 \mid \mathbf{x}_i)] \\
&= \mathbb{P}(y_i = 1 \mid \mathbf{x}_i)\, [1 - \mathbb{P}(y_i = 1 \mid \mathbf{x}_i)],
\end{aligned}$$

其中，第三个等号利用了离散随机变量条件期望的定义。我们将在本章第四节放松同方差假设进而讨论最小二乘估计量的大样本性质。

最小二乘估计

接下来我们讨论回归系数 $\boldsymbol{\beta}$ 的估计问题。由假设 2.1 和 2.2，我们知道总体回归函数与线性投影等价，估计 $\boldsymbol{\beta}$，实际上是对线性投影系数进行估计。由定义1.3，线性投影系数最小化如下均方预测误差：

$$S(\boldsymbol{b}) = \mathbb{E}\left((\mathrm{y} - \mathbf{x}'\boldsymbol{b})^2\right).$$

因此，给定样本 $(y_i, \mathbf{x}_i)_{i=1}^n$，我们定义第 i 个观测值的残差（residual）：

$$y_i - \mathbf{x}_i'\boldsymbol{b},$$

$\boldsymbol{\beta}$ 的估计量可以通过最小化残差平方和（residual sum of squares, RSS）得到。这实际上利用了数理统计中的矩估计思想，即，利用一阶样本矩（样本均值）来估计总体期望中相应的参数。因为样本量 n 固定，最小化残差平方的样本均值等价于最小化残差平方和。

定义 2.2. 普通最小二乘法（Ordinary Least Squares, OLS）

给定样本 $(y_i, \mathbf{x}_i)_{i=1}^n$，最小二乘法通过最小化如下残差平方和估计多元线性回归模型的系数 $\boldsymbol{\beta}$：

$$RSS(\boldsymbol{b}) \equiv \sum_{i=1}^n (y_i - \mathbf{x}_i'\boldsymbol{b})^2 = (\mathbf{y} - \mathbf{X}\boldsymbol{b})'(\mathbf{y} - \mathbf{X}\boldsymbol{b}).$$

记 $\boldsymbol{\beta}$ 的最小二乘估计量为 $\hat{\boldsymbol{\beta}}_{ols}$，我们有：

$$\hat{\boldsymbol{\beta}}_{ols} = \underset{\boldsymbol{b} \in \mathbb{R}^K}{\arg\min}\, RSS(\boldsymbol{b}).$$

♣

定义2.2中 $\boldsymbol{b}$，$\boldsymbol{\beta}$ 和 $\hat{\boldsymbol{\beta}}_{ols}$ 之间的关系见图 2.1。最小二乘估计量的解析解可以通过最优化的一阶条件 $\partial RSS(\boldsymbol{b})/\partial \boldsymbol{b} = \mathbf{0}$ 求得。注意到：

$$\begin{aligned}
RSS(\boldsymbol{b}) &= (\mathbf{y} - \mathbf{X}\boldsymbol{b})'(\mathbf{y} - \mathbf{X}\boldsymbol{b}) = (\mathbf{y}' - \boldsymbol{b}'\mathbf{X}')(\mathbf{y} - \mathbf{X}\boldsymbol{b}) \\
&= \mathbf{y}'\mathbf{y} - 2\mathbf{y}'\mathbf{X}\boldsymbol{b} + \boldsymbol{b}'\mathbf{X}'\mathbf{X}\boldsymbol{b} \qquad \text{（因为标量 } \mathbf{y}'\mathbf{X}\boldsymbol{b} \text{ 等于 } \boldsymbol{b}'\mathbf{X}'\mathbf{y}\text{）}
\end{aligned}$$

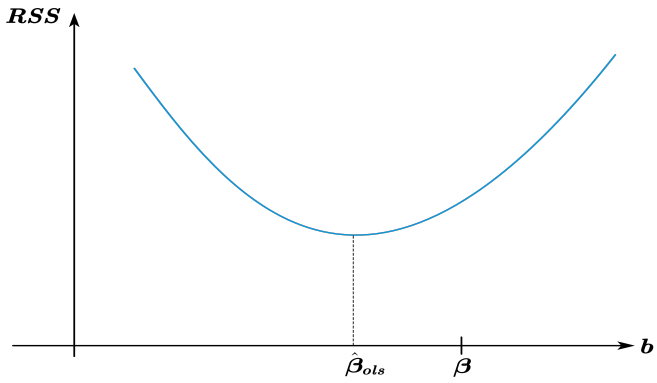

图 2.1: 最小二乘估计

$$\equiv \mathbf{y}'\mathbf{y} - 2\boldsymbol{a}'\boldsymbol{b} + \boldsymbol{b}'\mathbf{A}\boldsymbol{b}. \qquad (\text{定义 } a = \mathbf{X}'\mathbf{y},\ \mathbf{A} = \mathbf{X}'\mathbf{X})$$

因此，由 (C.5)，一阶条件可以表示为：

$$\frac{\partial RSS(\boldsymbol{b})}{\partial \boldsymbol{b}}\Big|_{\boldsymbol{b}=\hat{\boldsymbol{\beta}}_{ols}} = -2\boldsymbol{a} + 2\mathbf{A}\hat{\boldsymbol{\beta}}_{ols} = \mathbf{0}.$$

上式经过整理可得到最小二乘估计量满足的正则方程：

$$\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}}_{ols} = \mathbf{X}'\mathbf{y}. \tag{2.7}$$

由假设 2.3，$\mathbf{X}'\mathbf{X}$ 是正定阵并且可逆[2]，于是，我们有：

$$\hat{\boldsymbol{\beta}}_{ols} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y}. \tag{2.8}$$

用最小二乘估计量 $\hat{\boldsymbol{\beta}}_{ols}$ 替换总体回归函数中的真实参数 $\boldsymbol{\beta}$ 便得到样本回归函数，第 i 个观测的样本回归函数的取值为 $\mathbf{x}_i'\hat{\boldsymbol{\beta}}_{ols}$，我们将其称为观测 i 的拟合值（fitted value）并记为 $\hat{y}_i$。所有样本观测的拟合值构成的向量为：

$$\hat{\mathbf{y}} \equiv \mathbf{X}\hat{\boldsymbol{\beta}}_{ols} = \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y}.$$

由正则方程 (2.7) 有：

$$\mathbf{X}'(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_{ols}) = \mathbf{X}'\boldsymbol{e} = \mathbf{0}, \tag{2.9}$$

其中，$\boldsymbol{e} = (e_1, \ldots, e_n)' = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_{ols} = \mathbf{y} - \hat{\mathbf{y}}$，称为残差。若回归中含有常数项，则 $\mathbf{X}'$ 的第一行是 n 个 1 构成的行向量。因此，由式 (2.9)，含有常数项的回归满足 $\sum_{i=1}^n e_i = 0$。由 $\mathbf{X}$ 的定义，式 (2.9) 可以等价地写成均值形式：

$$\frac{1}{n}\sum_{i=1}^n \mathbf{x}_i e_i = \mathbf{0}. \tag{2.10}$$

在定理 1.7 中，线性投影误差 e 满足 $\mathbb{E}(\mathbf{x}e) = \mathbf{0}$，式 (2.10) 是此正交性质的样本形式。我们把 (2.9) 式称为残差与解释变量正交。

图 2.2 说明了最小二乘法的几何意义。此图中有两个解释变量 $\mathbf{x}_1$ 和 $\mathbf{x}_2$，$\mathbf{y}$ 对这两个变量做回归的几何意义是 $\mathbf{y}$ 对由这两个变量张成的平面做投影，落在平面上的投影部分

[2] 因为 $\mathbf{X}'\mathbf{X}$ 正定，所以目标函数的二阶海塞矩阵正定，所以 $\hat{\boldsymbol{\beta}}_{ols}$ 是最小化的解。

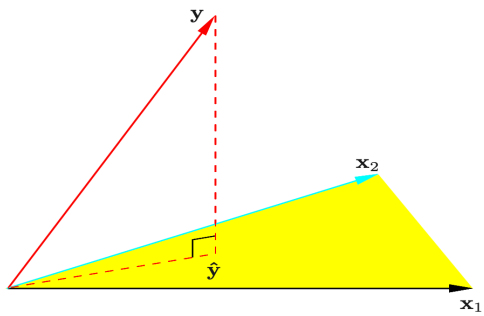

图 2.2: 最小二乘法的几何意义（图片来源：Hastie et al. (2016) 中图 3.2）

即为拟合值 $\hat{\mathbf{y}}$，$\mathbf{y}$ 到平面的垂线即为残差 $\boldsymbol{e}$。因为：

$$\hat{\mathbf{y}} = \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y},$$

所以，左乘矩阵 $\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'$ 起到对 $\mathbf{X}$ 的平面做投影的作用，我们将其记为 $\mathbf{P_X}$ 并称为投影矩阵。类似地，对于残差向量 $\boldsymbol{e}$，我们有：

$$\boldsymbol{e} = \mathbf{y} - \hat{\mathbf{y}} = \mathbf{y} - \mathbf{P_X}\cdot\mathbf{y} = \left(\mathbf{I}_n - \mathbf{P_X}\right)\mathbf{y}.$$

左乘矩阵 $\mathbf{I}_n - \mathbf{P_X}$ 的作用是得到对 $\mathbf{X}$ 的平面做投影的残差，我们将其记为 $\mathbf{M_X}$ 并称为消灭矩阵（annihilator）。容易验证 $\mathbf{P_X}$ 和 $\mathbf{M_X}$ 都是对称幂等阵，并且根据两个矩阵的几何意义容易知道：

$$\mathbf{P_X}\mathbf{X} = \mathbf{X}, \quad \mathbf{M_X}\mathbf{X} = \mathbf{0}.$$

另外，利用上面两个矩阵的性质，我们可以得到式 (1.7) 的样本对应形式，即，Frisch-Waugh 定理：

定理 2.1. Frisch-Waugh 定理

考虑如下回归形式：

$$\mathbf{y} = \begin{bmatrix} \mathbf{X}_r & \mathbf{X}_s \end{bmatrix}\begin{bmatrix} \boldsymbol{\beta}_r \\ \boldsymbol{\beta}_s \end{bmatrix} + \boldsymbol{\varepsilon},$$

其中，数据矩阵 $\mathbf{X}$ 被分解成 $\begin{bmatrix} \mathbf{X}_r & \mathbf{X}_s \end{bmatrix}$，回归系数 $\boldsymbol{\beta}$ 也做相应的分解。为了分别得到 $\boldsymbol{\beta}_r$ 和 $\boldsymbol{\beta}_s$ 的估计量，可以按照以下步骤操作：

1. 将 $\mathbf{y}$ 对 $\mathbf{X}_r$ 回归，得到残差向量 $\mathbf{y}^*$；
2. 将 $\mathbf{X}_s$ 的每列分别对 $\mathbf{X}_r$ 回归，所得残差向量组成矩阵 $\mathbf{X}_s^*$；
3. $\mathbf{y}^*$ 对 $\mathbf{X}_s^*$ 回归，得到 $\boldsymbol{\beta}_s$ 的 OLS 估计量 $\hat{\boldsymbol{\beta}}_s$；
4. 以 $\mathbf{y} - \mathbf{X}_s\hat{\boldsymbol{\beta}}_s$ 为被解释变量对 $\mathbf{X}_r$ 进行回归，得到 $\boldsymbol{\beta}_r$ 的 OLS 估计量 $\hat{\boldsymbol{\beta}}_r$。 ♡

证明 由最小二乘公式 (2.8) 和分块矩阵求逆公式（见式 (A.13)），我们有：

$$\begin{aligned}
\begin{bmatrix} \hat{\boldsymbol{\beta}}_r \\ \hat{\boldsymbol{\beta}}_s \end{bmatrix} &= \begin{bmatrix} \mathbf{X}_r'\mathbf{X}_r & \mathbf{X}_r'\mathbf{X}_s \\ \mathbf{X}_s'\mathbf{X}_r & \mathbf{X}_s'\mathbf{X}_s \end{bmatrix}^{-1}\begin{bmatrix} \mathbf{X}_r' \\ \mathbf{X}_s' \end{bmatrix}\mathbf{y} \\
&= \begin{bmatrix} \boldsymbol{A}_{11}^{-1} + \boldsymbol{A}_{11}^{-1}\boldsymbol{A}_{12}\boldsymbol{F}_2\boldsymbol{A}_{21}\boldsymbol{A}_{11}^{-1} & -\boldsymbol{A}_{11}^{-1}\boldsymbol{A}_{12}\boldsymbol{F}_2 \\ -\boldsymbol{F}_2\boldsymbol{A}_{21}\boldsymbol{A}_{11}^{-1} & \boldsymbol{F}_2 \end{bmatrix}\begin{bmatrix} \mathbf{X}_r'\mathbf{y} \\ \mathbf{X}_s'\mathbf{y} \end{bmatrix}
\end{aligned}$$

其中，

$$\begin{aligned}
\boldsymbol{A}_{11} &= \mathbf{X}_r'\mathbf{X}_r, \quad \boldsymbol{A}_{12} = \mathbf{X}_r'\mathbf{X}_s, \quad \boldsymbol{A}_{21} = \mathbf{X}_s'\mathbf{X}_r, \quad \boldsymbol{A}_{22} = \mathbf{X}_s'\mathbf{X}_s, \\
\boldsymbol{F}_2 &= \left(\boldsymbol{A}_{22} - \boldsymbol{A}_{21}\boldsymbol{A}_{11}^{-1}\boldsymbol{A}_{12}\right)^{-1} = \left(\mathbf{X}_s'\mathbf{M}_r\mathbf{X}_s\right)^{-1}, \quad (\mathbf{M}_r \equiv \mathbf{X}_r(\mathbf{X}_r'\mathbf{X}_r)^{-1}\mathbf{X}_r')
\end{aligned}$$

于是，

$$\begin{aligned}
\hat{\boldsymbol{\beta}}_s &= -\boldsymbol{F}_2\boldsymbol{A}_{21}\boldsymbol{A}_{11}^{-1}\mathbf{X}_r'\mathbf{y} + \boldsymbol{F}_2\mathbf{X}_s'\mathbf{y} = \boldsymbol{F}_2\mathbf{X}_s'\mathbf{M}_r\mathbf{y} \\
&= \left(\mathbf{X}_s'\mathbf{M}_r\mathbf{X}_s\right)^{-1}\mathbf{X}_s'\mathbf{M}_r\mathbf{y} = \left[(\mathbf{M}_r\mathbf{X}_s)'(\mathbf{M}_r\mathbf{X}_s)\right]^{-1}(\mathbf{M}_r\mathbf{X}_s)'(\mathbf{M}_r\mathbf{y}),
\end{aligned}$$

其中，最后一个等号我们利用了 $\mathbf{M}_r$ 的幂等性质。由 $\mathbf{M}_r$ 的定义，我们知道 $\mathbf{M}_r\mathbf{X}_s = \mathbf{X}_s^*$，$\mathbf{M}_r\mathbf{y} = \mathbf{y}^*$。于是，

$$\hat{\boldsymbol{\beta}}_s = \left(\mathbf{X}_s^{*\prime}\mathbf{X}_s^*\right)^{-1}\mathbf{X}_s^{*\prime}\mathbf{y}^*.$$

定理的步骤 1-3 得证。事实上，由 $\mathbf{M}_r$ 的幂等性，可以直接用 $\mathbf{y}$ 对 $\mathbf{X}_s^*$ 回归得到 $\hat{\boldsymbol{\beta}}_s$。另一方面，由式 (2.7)，

$$\mathbf{X}_r'\mathbf{X}_r\hat{\boldsymbol{\beta}}_r + \mathbf{X}_r'\mathbf{X}_s\hat{\boldsymbol{\beta}}_s = \mathbf{X}_r'\mathbf{y} \quad \Rightarrow \quad \hat{\boldsymbol{\beta}}_r = \left(\mathbf{X}_r'\mathbf{X}_r\right)^{-1}\mathbf{X}_r(\mathbf{y} - \mathbf{X}_s\hat{\boldsymbol{\beta}}_s).$$

□

图 2.3 说明了 Frisch-Waugh 定理的几何意义。在图中，$\mathbf{x}_2$ 对 $\mathbf{x}_1$ 做回归得到残差 $\mathbf{z}$，$\mathbf{y}$ 对 $\mathbf{z}$ 回归可以得到 $\mathbf{x}_2$ 在如下多元线性回归中的系数 β_2 的估计量：

$$\mathbf{y} = \beta_1\mathbf{x}_1 + \beta_2\mathbf{x}_2 + \varepsilon.$$

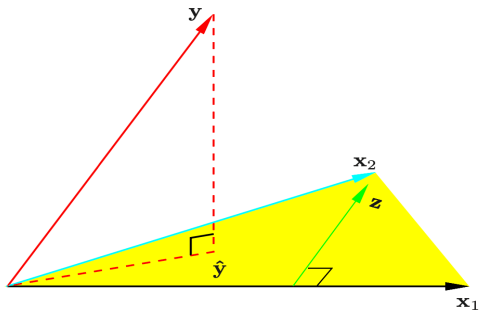

图 2.3: Frisch-Waugh 定理（图片来源：Hastie et al. (2016) 中图 3.4）

基于公式 (2.8) 和 Frisch-Waugh 定理，我们可以推导出如下常见回归的系数表达式：

- 只含有常数项的线性回归：$y_i = \beta_0 + \varepsilon_i$，$i = 1, 2, \ldots, n$，

$$\hat{\beta}_{0,ols} = \bar{y}, \quad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i. \tag{2.11}$$

- 简单线性回归：$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$，$i = 1, 2, \ldots, n$，

$$\hat{\beta}_{1,ols} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \quad \hat{\beta}_{0,ols} = \bar{y} - \hat{\beta}_{1,ols}\bar{x}, \quad \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i. \tag{2.12}$$

  简单线性回归的系数估计量实际上是式 (1.6) 的样本形式。

- 含两个解释变量的线性回归：$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i$，$i = 1, 2, \ldots, n$，

$$\hat{\beta}_{1,ols} = \frac{\sum_{i=1}^{n}\hat{r}_{i1}y_i}{\sum_{i=1}^{n}\hat{r}_{i1}^2}, \tag{2.13}$$

其中，$\hat{r}_{i1}$ 是 x_{i1} 对 x_{i2} 和常数项做回归得到的残差。

Stata 的非官方命令 reganat 提供了 Frisch-Waugh 定理的可视化功能，可以使用如下命令安装：

```
. ssc install reganat, replace
```

2.2 拟合优度与小样本性质

拟合优度

在本节中，我们首先讨论线性回归模型的拟合优度（goodness of fit）问题。所谓拟合优度是指样本回归函数对样本观测值的拟合程度。

第一种可以用来衡量拟合优度的指标是非中心化 R^2，定义如下：

定义 2.3. 非中心化 R^2

非中心化 R^2 的定义为：

$$R_{uc}^2 \equiv 1 - \frac{\boldsymbol{e}'\boldsymbol{e}}{\mathbf{y}'\mathbf{y}}.$$

它衡量的是被解释变量的变化由解释变量解释的比例。 ♣

被解释变量 $\mathbf{y}$ 的变化可以用 $\sum_{i=1}^n y_i^2 = \mathbf{y}'\mathbf{y}$ 度量。根据 $\boldsymbol{e} = \mathbf{y} - \hat{\mathbf{y}}$，容易证明：

$$\begin{aligned}
\mathbf{y}'\mathbf{y} &= (\hat{\mathbf{y}} + \boldsymbol{e})'(\hat{\mathbf{y}} + \boldsymbol{e}) = \hat{\mathbf{y}}'\hat{\mathbf{y}} + 2\hat{\mathbf{y}}'\boldsymbol{e} + \boldsymbol{e}'\boldsymbol{e} && (\hat{\mathbf{y}}'\boldsymbol{e} \text{ 和 } \boldsymbol{e}'\hat{\mathbf{y}} \text{ 是标量}) \\
&= \hat{\mathbf{y}}'\hat{\mathbf{y}} + 2\hat{\boldsymbol{\beta}}_{ols}'\mathbf{X}'\boldsymbol{e} + \boldsymbol{e}'\boldsymbol{e} && (\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}}_{ols}) \\
&= \hat{\mathbf{y}}'\hat{\mathbf{y}} + \boldsymbol{e}'\boldsymbol{e}. && (\mathbf{X}'\boldsymbol{e} = \mathbf{0})
\end{aligned}$$

所以，

$$R_{uc}^2 = \frac{\hat{\mathbf{y}}'\hat{\mathbf{y}}}{\mathbf{y}'\mathbf{y}}.$$

上式分子衡量的是拟合值的变化，是 $\mathbf{y}$ 的变化中可以由解释变量解释的部分。

非中心化 R^2 关注的是所有解释变量对 $\mathbf{y}$ 的解释能力，这其中也包含常数项。但是，线性回归中通常都要包含常数项。所以，我们更关注除常数项之外的解释变量对 $\mathbf{y}$ 的解释能力。为此，我们定义第二种衡量拟合优度的指标：（中心化）R^2，也称为可决系数（coefficient of determination）。

定义 2.4. （中心化）R^2

（中心化）R^2，也称为可决系数，衡量的是除常数项之外的解释变量对 $\mathbf{y}$ 的解释能力，其定义如下：

$$R^2 \equiv 1 - \frac{\sum_{i=1}^n e_i^2}{\sum_{i=1}^n (y_i - \bar{y})^2}. \tag{2.14}$$

♣

为了更好地理解 R^2 的定义，我们首先证明：如果线性回归中含有常数项，则有：

$$\sum_{i=1}^{n}(y_i-\bar{y})^2=\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2+\sum_{i=1}^{n}e_i^2. \tag{2.15}$$

证明 将等式左边写成向量内积形式：

$$\sum_{i=1}^{n}(y_i-\bar{y})^2=(\mathbf{y}-\mathbf{1}_n\bar{y})'(\mathbf{y}-\mathbf{1}_n\bar{y}),$$

其中，$\mathbf{1}_n$ 表示 $n\times 1$ 维由 1 构成的列向量。注意到：

$$\begin{aligned}
&(\mathbf{y}-\mathbf{1}_n\bar{y})'(\mathbf{y}-\mathbf{1}_n\bar{y})\\
=\ &(\mathbf{y}-\hat{\mathbf{y}}+\hat{\mathbf{y}}-\mathbf{1}_n\bar{y})'(\mathbf{y}-\hat{\mathbf{y}}+\hat{\mathbf{y}}-\mathbf{1}_n\bar{y})\\
=\ &(\hat{\mathbf{y}}-\mathbf{1}_n\bar{y})'(\hat{\mathbf{y}}-\mathbf{1}_n\bar{y})+(\mathbf{y}-\hat{\mathbf{y}})'(\mathbf{y}-\hat{\mathbf{y}})+2(\hat{\mathbf{y}}-\mathbf{1}_n\bar{y})'\underbrace{(\mathbf{y}-\hat{\mathbf{y}})}_{\boldsymbol{e}}.
\end{aligned}$$

若回归中含有常数项，根据式 (2.9)，则有：$\mathbf{1}_n'\boldsymbol{e}=0$，$\hat{\mathbf{y}}'\boldsymbol{e}=\hat{\boldsymbol{\beta}}_{ols}'\mathbf{X}'\boldsymbol{e}=0$，式 (2.15) 得证。 □

在式 (2.15) 中，$\sum_{i=1}^{n}(y_i-\bar{y})^2$ 称为总平方和（total sum of squares, TSS），$\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2$ 称为回归平方和（explained sum of squares, ESS）。因此，式 (2.15) 的结论也可以表示为：TSS = ESS + RSS。由式 (2.15)，R^2 可以表示成：

$$R^2=\frac{\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2}.$$

根据式 (2.11)，$\bar{y}$ 是只含有常数项的线性回归的拟合值，因此，R^2 衡量的是除常数项之外的解释变量对 $\mathbf{y}$ 的解释能力。若回归中不含常数项，使用式 (2.14) 计算 R^2 可能会得到取值为负的结果。事实上，

$$1-\frac{\sum_{i=1}^{n}e_i^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2}<0$$

可以推出

$$\sum_{i=1}^{n}e_i^2>\sum_{i=1}^{n}(y_i-\bar{y})^2.$$

这表明具有负值 R^2 的回归的解释变量对 $\mathbf{y}$ 的拟合能力不及样本均值对 $\mathbf{y}$ 的拟合能力。在线性回归模型中不包含常数项的情况下，Stata 会汇报非中心化 R^2。

R^2 的一个重要性质是如果总体回归函数中增加解释变量，即使增加的解释变量与被解释变量无关，R^2 也只增不减。Greene (2018) 中的定理 3.6 给出了此性质的详细证明。我们也可以从最小二乘估计的原理出发来理解这一性质。考虑如下两个回归：

$$\begin{aligned}
y_i &= \beta_0+\beta_1x_{i1}+\varepsilon_i,\\
y_i &= \beta_0+\beta_1x_{i1}+\beta_2x_{i2}+\varepsilon_i.
\end{aligned}$$

第一个回归可以看成是第二个回归的约束形式，即，施加了约束 $\beta_2=0$。由定义2.2，第一个回归的残差平方和是通过一个有约束的最小化问题得到的，其结果一定不小于第二个回归的残差平方和。因此，由 R^2 的定义，第二个回归的 R^2 不小于第一个回归的 R^2。如果用 R^2 作为模型选择的标准，它只会选择最复杂的模型。但是建模的原则是：在保证

模型能完成其既定任务的前提下，模型要尽可能的简洁。因此，我们有如下考虑了模型复杂度的拟合优度指标：[3]

定义 2.5. 调整的 R^2（adjusted R^2）

调整的 R^2 定义如下：

$$\bar{R}^2 \equiv 1 - \frac{\sum_{i=1}^n e_i^2/(n-K)}{\sum_{i=1}^n (y_i - \bar{y})^2/(n-1)}. \tag{2.16}$$

调整的 R^2 通常也记为 $adj\text{-}R^2$。 ♣

$adj\text{-}R^2$ 的调整实际上是对模型自由度的调整，是对增加解释变量施加了惩罚。模型复杂度越高，解释变量个数越多，在样本量固定的情况下，必然引起模型自由度的减少。$adj\text{-}R^2$ 将 RSS 和 TSS 分别除以各自的自由度 $n-K$ 和 $n-1$，可能使得具有更多的解释变量的模型的 $adj\text{-}R^2$ 减小。$adj\text{-}R^2$ 的一个缺点是它的取值可能为负。

虽然上面定义的几种拟合优度指标衡量了解释变量的拟合能力，但是，若回归分析的目的是为了揭示解释变量和被解释变量之间的因果关系，则我们不用过度关注拟合优度。具有很低的 R^2 的线性回归模型中的系数估计量依然可能是解释变量对被解释变量因果效应的合理估计。这一点我们在第 **??** 章中再进一步说明。

小样本性质

接下来，我们分析最小二乘估计量的有限样本性质。有限样本性质是指对任意给定的 n $(\geqslant K)$ 都成立的性质。这些性质的成立是因为我们施加了诸如严格外生和球形扰动等较强的假设。

定理 2.2. 最小二乘估计量的小样本性质

(a) （无偏性）若假设 2.1-2.3 成立，则：$\mathbb{E}(\hat{\boldsymbol{\beta}}_{ols} \mid \mathbf{X}) = \boldsymbol{\beta}$；

(b) 若假设 2.1-2.4 成立，则：$\mathrm{Var}(\hat{\boldsymbol{\beta}}_{ols} \mid \mathbf{X}) = \sigma^2 \cdot (\mathbf{X}'\mathbf{X})^{-1}$；

(c) （Gauss-Markov 定理）若假设 2.1-2.4 成立，则：最小二乘估计量是所有线性无偏估计量中最有效的，即，对任意的关于 $\mathbf{y}$ 线性并且无偏的估计量 $\hat{\boldsymbol{\beta}}$，$\mathrm{Var}(\hat{\boldsymbol{\beta}} \mid \mathbf{X}) - \mathrm{Var}(\hat{\boldsymbol{\beta}}_{ols} \mid \mathbf{X}) \succeq \mathbf{0}$；[a]

(d) 若假设 2.1-2.4 成立，$\mathrm{Cov}(\hat{\boldsymbol{\beta}}_{ols}, \boldsymbol{e} \mid \mathbf{X}) = \mathbf{0}$，其中 $\boldsymbol{e} \equiv \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_{ols}$。

[a] 因此，最小二乘估计量被称为最优线性无偏估计量（Best Linear Unbiased Estimator, BLUE）。 ♡

证明 (a) 我们证明 $\mathbb{E}(\hat{\boldsymbol{\beta}}_{ols} - \boldsymbol{\beta} \mid \mathbf{X}) = \mathbf{0}$。注意到：

$$\hat{\boldsymbol{\beta}}_{ols} - \boldsymbol{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon} \equiv \mathbf{A}\boldsymbol{\varepsilon}. \quad (\mathbf{A} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}') \tag{2.17}$$

因此，

$$\mathbb{E}(\hat{\boldsymbol{\beta}}_{ols} - \boldsymbol{\beta} \mid \mathbf{X}) = \mathbb{E}(\mathbf{A}\boldsymbol{\varepsilon}|\mathbf{X}) = \mathbf{A}\mathbb{E}(\boldsymbol{\varepsilon}|\mathbf{X}) = \mathbf{0}.$$

(b) 对于条件方差，我们有：

$\mathrm{Var}(\hat{\boldsymbol{\beta}}_{ols} \mid \mathbf{X}) = \mathrm{Var}(\hat{\boldsymbol{\beta}}_{ols} - \boldsymbol{\beta} \mid \mathbf{X})$

[3] Greene (2018) 中的定理 3.7 详细讨论了调整的 R^2 的性质。

$$
\begin{aligned}
&= \mathrm{Var}(\mathbf{A}\boldsymbol{\varepsilon}|\mathbf{X}) && (\text{由 (a) 中 } \hat{\boldsymbol{\beta}}_{ols}-\boldsymbol{\beta} \text{ 的表达式}) \\
&= \mathbf{A}\mathbb{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'|\mathbf{X})\mathbf{A}' && (\text{由 (B.6) 和式 (2.6)}) \\
&= \sigma^2\mathbf{A}\mathbf{A}' && (\text{由式 (2.6)}) \\
&= \sigma^2\cdot(\mathbf{X}'\mathbf{X})^{-1}. && (\text{将 } \mathbf{A}=(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \text{ 代入})
\end{aligned}
$$

(c) 因为 $\hat{\boldsymbol{\beta}}$ 关于 $\mathbf{y}$ 是线性的，它可以记成 $\hat{\boldsymbol{\beta}}=\mathbf{C}\mathbf{y}$，其中 $\mathbf{C}$ 是某一由 $\mathbf{X}$ 确定的矩阵。令 $\mathbf{D}\equiv\mathbf{C}-\mathbf{A}$，其中 $\mathbf{A}\equiv(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$，则有：

$$
\hat{\boldsymbol{\beta}}=(\mathbf{D}+\mathbf{A})\mathbf{y}=\mathbf{D}\mathbf{X}\boldsymbol{\beta}+\mathbf{D}\boldsymbol{\varepsilon}+\hat{\boldsymbol{\beta}}_{ols}.
$$

上式两端同时取条件期望得到：

$$
\underbrace{\mathbb{E}(\hat{\boldsymbol{\beta}}|\mathbf{X})}_{\boldsymbol{\beta}}=\mathbf{D}\mathbf{X}\boldsymbol{\beta}+\underbrace{\mathbb{E}(\mathbf{D}\boldsymbol{\varepsilon}|\mathbf{X})}_{\mathbf{0}}+\underbrace{\mathbb{E}(\hat{\boldsymbol{\beta}}_{ols}|\mathbf{X})}_{\boldsymbol{\beta}}.
$$

因此，$\mathbf{D}\mathbf{X}\boldsymbol{\beta}=\mathbf{0}$。若对任意的 $\boldsymbol{\beta}$ 此结论都成立，必有 $\mathbf{D}\mathbf{X}=\mathbf{0}$。于是，$\hat{\boldsymbol{\beta}}=\mathbf{D}\boldsymbol{\varepsilon}+\hat{\boldsymbol{\beta}}_{ols}$，并且

$$
\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}=\mathbf{D}\boldsymbol{\varepsilon}+(\hat{\boldsymbol{\beta}}_{ols}-\boldsymbol{\beta})=(\mathbf{D}+\mathbf{A})\boldsymbol{\varepsilon}.
$$

因此，

$$
\begin{aligned}
\mathrm{Var}(\hat{\boldsymbol{\beta}}\mid\mathbf{X}) &= \mathrm{Var}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\mid\mathbf{X}) \\
&= \mathrm{Var}\left[(\mathbf{D}+\mathbf{A})\boldsymbol{\varepsilon}\mid\mathbf{X}\right] \\
&= (\mathbf{D}+\mathbf{A})\mathbb{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'|\mathbf{X})(\mathbf{D}'+\mathbf{A}') \\
&= \sigma^2\cdot(\mathbf{D}\mathbf{D}'+\mathbf{A}\mathbf{D}'+\mathbf{D}\mathbf{A}'+\mathbf{A}\mathbf{A}') \\
&= \sigma^2\cdot(\mathbf{D}\mathbf{D}'+\mathbf{A}\mathbf{A}') \\
&\geqslant \sigma^2\cdot(\mathbf{X}'\mathbf{X})^{-1}=\mathrm{Var}(\hat{\boldsymbol{\beta}}_{ols}\mid\mathbf{X}),
\end{aligned}
$$

其中，第四个等号成立是根据球形扰动假设；第五个等号用到了 $\mathbf{D}\mathbf{A}'=\mathbf{D}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}=\mathbf{0}$，同理，$\mathbf{A}\mathbf{D}'=\mathbf{0}$；不等号成立是由于 $\mathbf{D}\mathbf{D}'$ 半正定。

(d) 首先，注意到残差向量 $\boldsymbol{e}$ 可表示为

$$
\boldsymbol{e}=\mathbf{M}_{\mathbf{X}}\cdot\mathbf{y}=\mathbf{M}_{\mathbf{X}}\cdot(\mathbf{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon})=\mathbf{M}_{\mathbf{X}}\mathbf{X}\boldsymbol{\beta}+\mathbf{M}_{\mathbf{X}}\boldsymbol{\varepsilon}=\mathbf{M}_{\mathbf{X}}\boldsymbol{\varepsilon}. \tag{2.18}
$$

同时，我们有：$\mathbb{E}(\boldsymbol{e}\mid\mathbf{X})=\mathbb{E}(\mathbf{M}_{\mathbf{X}}\boldsymbol{\varepsilon}\mid\mathbf{X})=\mathbf{M}_{\mathbf{X}}\mathbb{E}(\boldsymbol{\varepsilon}\mid\mathbf{X})=\mathbf{0}$。

$$
\begin{aligned}
\mathrm{Cov}(\hat{\boldsymbol{\beta}}_{ols},\boldsymbol{e}\mid\mathbf{X}) &\equiv \mathbb{E}\left\{\left[\hat{\boldsymbol{\beta}}_{ols}-\mathbb{E}(\hat{\boldsymbol{\beta}}_{ols}\mid\mathbf{X})\right]\cdot[\boldsymbol{e}-\mathbb{E}(\boldsymbol{e}\mid\mathbf{X})]'\mid\mathbf{X}\right\} \\
&= \mathbb{E}\left\{\left[\hat{\boldsymbol{\beta}}_{ols}-\boldsymbol{\beta}\right]\cdot[\mathbf{M}_{\mathbf{X}}\boldsymbol{\varepsilon}]'\mid\mathbf{X}\right\} \\
&= \mathbb{E}\left\{\mathbf{A}\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\mathbf{M}_{\mathbf{X}}\mid\mathbf{X}\right\} \\
&= \mathbf{A}\mathbb{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\mid\mathbf{X})\mathbf{M}_{\mathbf{X}} \\
&= \sigma^2\mathbf{A}\mathbf{M}_{\mathbf{X}} \\
&= \mathbf{0},
\end{aligned}
$$

其中，最后一个等号用到 $\mathbf{A}\mathbf{M}_{\mathbf{X}}=(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{M}_{\mathbf{X}}=\mathbf{0}$。 □

Gauss-Markov 定理的结论是 $\mathrm{Var}(\hat{\boldsymbol{\beta}}\mid\mathbf{X})-\mathrm{Var}(\hat{\boldsymbol{\beta}}_{ols}\mid\mathbf{X})\succeq\mathbf{0}$，即，$\mathrm{Var}(\hat{\boldsymbol{\beta}}\mid\mathbf{X})-$

$\mathrm{Var}(\hat{\boldsymbol{\beta}}_{ols} \mid \mathbf{X})$ 是半正定矩阵。为了理解此结论，我们令 $\boldsymbol{a}_k$ 表示 $K \times 1$ 维的列向量，它的第 k 行元素为 1，其余行的元素为 0。由 Gauss-Markov 定理的结论，我们有：

$$\begin{aligned} & \boldsymbol{a}_k' \left\{ \mathrm{Var}(\hat{\boldsymbol{\beta}} \mid \mathbf{X}) - \mathrm{Var}(\hat{\boldsymbol{\beta}}_{ols} \mid \mathbf{X}) \right\} \boldsymbol{a}_k \\ = \; & \boldsymbol{a}_k' \mathrm{Var}(\hat{\boldsymbol{\beta}} \mid \mathbf{X}) \boldsymbol{a}_k - \boldsymbol{a}_k' \mathrm{Var}(\hat{\boldsymbol{\beta}}_{ols} \mid \mathbf{X}) \boldsymbol{a}_k \\ \geqslant \; & 0. \end{aligned}$$

因为，一个矩阵左乘 $\boldsymbol{a}_k'$ 的同时再右乘 $\boldsymbol{a}_k$ 的作用是取出其主对角线上的第 k 个元素，所以，$\hat{\boldsymbol{\beta}}_k$ 的方差不小于 $\hat{\boldsymbol{\beta}}_{k,ols}$ 的方差。

单个回归系数估计量的方差

定理 2.2 中结论 (b) 给出了最小二乘估计量的条件方差，单个系数估计量的条件方差是矩阵 $\sigma^2 \cdot (\mathbf{X}'\mathbf{X})^{-1}$ 主对角线上的元素。然而，推导单个系数估计量的条件方差的解析式对于深入理解最小二乘估计量的条件方差具有重要的作用。

我们考虑如下回归中系数 β_1 的最小二乘估计量的条件方差：

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_K x_{iK} + \varepsilon_i.$$

事实上，由 Frisch-Waugh 定理，式 (2.13) 可以推广为上式中 β_1 的最小二乘估计量表达式：

$$\hat{\beta}_{1,ols} = \frac{\sum_{i=1}^n \hat{r}_{i1} y_i}{\sum_{i=1}^n \hat{r}_{i1}^2}, \tag{2.19}$$

其中，$\hat{r}_{i1}$ 是 x_{i1} 对 $x_{i2}, \ldots, x_{iK}$ 和常数项做回归得到的残差。基于上式，我们有如下定理：

定理 2.3. 单个系数估计量的方差

若假设 2.1-2.4 成立，我们有：

$$\mathrm{Var}(\hat{\beta}_{j,ols} \mid \mathbf{X}) = \frac{\sigma^2}{TSS_j(1 - R_j^2)}, \quad j = 1, \ldots, K, \tag{2.20}$$

其中，$TSS_j = \sum_{i=1}^n (x_{ij} - \bar{x}_j)^2$，$\bar{x}_j = \sum_{i=1}^n x_{ij}/n$，$R_j^2$ 是 x_{ij} 对所有其他解释变量和常数项做回归得到的（中心化）R^2。

证明 不失一般性，我们对 $j = 1$ 的情况进行证明。由式 (2.19)，我们有：

$$\begin{aligned} \hat{\beta}_{1,ols} &= \frac{\sum_{i=1}^n \hat{r}_{i1} y_i}{\sum_{i=1}^n \hat{r}_{i1}^2} \\ &= \frac{\sum_{i=1}^n \hat{r}_{i1} (\beta_0 + \beta_1 x_{i1} + \cdots + \beta_K x_{iK} + \varepsilon_i)}{\sum_{i=1}^n \hat{r}_{i1}^2} \\ &= \beta_1 \frac{\sum_{i=1}^n \hat{r}_{i1} x_{i1}}{\sum_{i=1}^n \hat{r}_{i1}^2} + \frac{\sum_{i=1}^n \hat{r}_{i1} \varepsilon_i}{\sum_{i=1}^n \hat{r}_{i1}^2} \\ &= \beta_1 + \frac{\sum_{i=1}^n \hat{r}_{i1} \varepsilon_i}{\sum_{i=1}^n \hat{r}_{i1}^2}, \end{aligned}$$

其中，第三个等号成立用到了残差与回归解释变量正交；第四个等号成立是因为

$$\sum_{i=1}^n \hat{r}_{i1} x_{i1} = \sum_{i=1}^n \hat{r}_{i1} (\hat{x}_{i1} + \hat{r}_{i1}) = \sum_{i=1}^n \hat{r}_{i1}^2,$$

其中，$\hat{x}_{i1}$ 是 x_{i1} 对所有其他解释变量和常数项做回归得到的拟合值，此拟合值与残差 $\hat{r}_{i1}$

正交。

因此，我们有：

$$
\begin{aligned}
\mathrm{Var}(\hat{\beta}_{1,ols} \mid \mathbf{X}) &= \mathrm{Var}(\hat{\beta}_{1,ols} - \beta_1 \mid \mathbf{X}) \\
&= \mathrm{Var}\left(\frac{\sum_{i=1}^n \hat{r}_{i1}\varepsilon_i}{\sum_{i=1}^n \hat{r}_{i1}^2} \,\middle|\, \mathbf{X} \right) \\
&= \frac{1}{(\sum_{i=1}^n \hat{r}_{i1}^2)^2} \sum_{i=1}^n \hat{r}_{i1}^2 \mathrm{Var}(\varepsilon_i \mid \mathbf{X}) && \text{（以 } \mathbf{X} \text{ 为条件 } \hat{r}_{i1} \text{ 可认为非随机）} \\
&= \frac{\sigma^2}{\sum_{i=1}^n \hat{r}_{i1}^2}. && \text{（由同方差假设）}
\end{aligned}
$$

同时，注意到 $\sum_{i=1}^n \hat{r}_{i1}^2$ 为 x_{i1} 对所有其他解释变量和常数项做回归得到的残差平方和，由式 (2.14)，$\sum_{i=1}^n \hat{r}_{i1}^2 = (1 - R_1^2)TSS_1$。 □

2.3 古典模型的假设检验

本节讨论小样本情形下的多元线性回归模型的假设检验问题。假设检验在计量实证分析中具有广泛的应用。例如，资本资产定价模型（Capital Asset Pricing Model, CAPM）是金融经济学中的重要模型，该模型指出：

$$\mathbb{E}\left[r_{jt} - r_f\right] = \beta_j \mathbb{E}\left[r_{mt} - r_f\right], \tag{2.21}$$

其中，r_{jt} 是 t 时刻风险资产 j 的收益率，r_{mt} 是 t 时刻市场组合的收益率，r_f 表示无风险资产的收益率。系统性风险（systematic risk）测度 β_j 满足：

$$\beta_j = \frac{\mathrm{Cov}(r_{jt}, r_{mt})}{\mathrm{Var}(r_{mt})}.$$

注意到 β_j 具有线性投影系数 (1.6) 的形式，因此可以用最小二乘法进行估计。为此，定义

$$u_{jt} = r_{jt} - \mathbb{E}(r_{jt}), \quad u_{mt} = r_{mt} - \mathbb{E}(r_{mt}).$$

于是，式 (2.21) 可以表示为：

$$r_{jt} - r_f = \beta_j(r_{mt} - r_f) + \varepsilon_{jt}, \quad \varepsilon_{jt} = u_{jt} - \beta_j u_{mt}.$$

我们考虑如下回归：

$$r_{jt} - r_f = \alpha_j + \beta_j(r_{mt} - r_f) + \varepsilon_{jt}, \quad \varepsilon_{jt} = u_{jt} - \beta_j u_{mt}.$$

若 CAPM 对资产 j 有效，则有 $\alpha_j = 0$，此条件是否成立可以通过假设[4]检验来分析。我们将想要检验的假设称为原假设（null hypothesis），通常记为 $\mathbb{H}_0$。在上述例子中，原假设为：

$$\mathbb{H}_0: \ \alpha_j = 0.$$

等式约束的 $\mathbb{H}_0$ 的否命题称为双边备择假设（alternative hypothesis），备择假设通常记为 $\mathbb{H}_1$。

[4] 一般地，“假设”是对模型参数或其函数的具体取值所作的陈述。

在 2.1 节的开始我们提到“模型”是对解释变量和被解释变量的联合分布施加的一组假设，这组假设通常被称为保留假设（maintained hypothesis）。假设检验需要在保留假设和原假设同时成立的情况下构造检验统计量并推导出它的理论分布。在多元线性回归的小样本假设检验中，保留假设是由 CLRM 的四条假设和 ε 的条件正态假设组成的。如果保留假设成立，我们称模型是正确设定的。即使原假设为真，也可能出现由于模型误设导致检验统计量不服从理论分布的情况。

为了推导小样本下的假设检验理论，我们施加如下正态假设：

假设 2.5：（误差项正态假设）

$$\varepsilon \mid \mathbf{X} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \boldsymbol{I}_n),$$

即，以 $\mathbf{X}$ 为条件，ε 服从联合正态分布。

正态分布的密度函数只和分布的期望和方差有关，一般地，以 $\mathbf{X}$ 为条件的条件正态分布的期望和方差都是 $\mathbf{X}$ 的函数。但是，假设 2.5 要求 ε 的条件正态分布的期望和方差都与 $\mathbf{X}$ 无关，因此，ε 的无条件分布也是 $\mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I}_n)$，并且 ε 与 $\mathbf{X}$ 相互独立。正态分布有一些良好的性质对于条件正态分布依然成立：

- 若两个随机变量以 $\mathbf{X}$ 为条件服从联合正态且不相关，则两个随机变量以 $\mathbf{X}$ 为条件相互独立；
- 若 ε 以 $\mathbf{X}$ 为条件服从正态分布，则 $\mathbf{A}\varepsilon$ 以 $\mathbf{X}$ 为条件也服从正态分布，其中 $\mathbf{A}$ 是 $\mathbf{X}$ 的函数。

于是，由式 (2.17)，在假设 2.1-2.5 成立的情况下，我们有：

$$(\hat{\boldsymbol{\beta}}_{ols} - \boldsymbol{\beta}) \mid \mathbf{X} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \cdot (\mathbf{X}'\mathbf{X})^{-1}). \tag{2.22}$$

又因为，$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \varepsilon$，我们可以得到 $\mathbf{y}$ 的条件分布为：

$$\mathbf{y} \mid \mathbf{X} \sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n).$$

因此，若 y_i 是离散变量，假设 2.5 不再成立。

单个回归系数的假设检验

我们首先考虑单个回归系数的假设检验问题。需要说明的是本节讨论的假设检验问题都是线性假设检验问题，即，原假设是关于回归系数的线性函数的一个约束条件。

例如，我们关心第 k 个解释变量的系数 β_k 是否等于某一常数 β_k^0，则原假设为：$\mathbb{H}_0: \beta_k = \beta_k^0$，备择假设为 $\mathbb{H}_1: \beta_k \neq \beta_k^0$。由式 (2.22)，在假设 2.1-2.5 和原假设成立的情况下，

$$(\hat{\beta}_{k,ols} - \beta_k^0) \mid \mathbf{X} \sim \mathcal{N}\left(0, \sigma^2 \cdot \left((\mathbf{X}'\mathbf{X})^{-1}\right)_{kk}\right),$$

其中，$\left((\mathbf{X}'\mathbf{X})^{-1}\right)_{kk}$ 是矩阵 $(\mathbf{X}'\mathbf{X})^{-1}$ 主对角线上的第 k 个元素。所以，若定义

$$z_k \equiv \frac{\hat{\beta}_{k,ols} - \beta_k^0}{\sqrt{\sigma^2 \cdot \left((\mathbf{X}'\mathbf{X})^{-1}\right)_{kk}}},$$

则 z_k 服从标准正态分布。然而，虽然 z_k 的分布已知，但是，将 z_k 作为检验统计量依然

是不可行的，因为 σ^2 是未知的。为了构造可行的检验统计量，我们先讨论 σ^2 如何估计。

定理 2.4. σ^2 的估计

σ^2 的估计量，记为 s^2，具有如下形式：

$$s^2 \equiv \frac{\boldsymbol{e}'\boldsymbol{e}}{n-K}.$$

只要 $n > K$，在假设 2.1-2.4 成立的情况下，我们有：$\mathbb{E}(s^2|\mathbf{X}) = \sigma^2$，因此，$s^2$ 是 σ^2 的无偏估计。 ♡

证明 因为 $s^2 = \boldsymbol{e}'\boldsymbol{e}/(n-K)$，我们证明：

$$\mathbb{E}(\boldsymbol{e}'\boldsymbol{e}|\mathbf{X}) = (n-K)\sigma^2.$$

根据式 (2.18)，上式的证明可以进一步拆分成如下两个性质的证明：

(1) $\mathbb{E}(\boldsymbol{e}'\boldsymbol{e}|\mathbf{X}) = \mathbb{E}(\boldsymbol{\varepsilon}'\mathbf{M}_\mathbf{X}\boldsymbol{\varepsilon}|\mathbf{X}) = \sigma^2 \cdot \mathrm{tr}(\mathbf{M}_\mathbf{X})$；

(2) $\mathrm{tr}(\mathbf{M}_\mathbf{X}) = n-K$。

首先证明 (1)，因为 $\boldsymbol{\varepsilon}'\mathbf{M}_\mathbf{X}\boldsymbol{\varepsilon} = \sum_{i=1}^n \sum_{j=1}^n m_{ij}\varepsilon_i\varepsilon_j$，其中，$m_{ij}$ 是 $\mathbf{M}_\mathbf{X}$ 的第 (i,j) 位置上的元素，于是，我们有：

$$\mathbb{E}(\boldsymbol{\varepsilon}'\mathbf{M}_\mathbf{X}\boldsymbol{\varepsilon}|\mathbf{X}) = \sum_{i=1}^n\sum_{j=1}^n m_{ij}\mathbb{E}(\varepsilon_i\varepsilon_j|\mathbf{X}) = \sigma^2\sum_{i=1}^n m_{ii} = \sigma^2 \cdot \mathrm{tr}(\mathbf{M}_\mathbf{X}),$$

其中，第一个等号成立是因为 m_{ij} 是由 $\mathbf{X}$ 决定的，根据定理 1.3，m_{ij} 可以提到期望符号的外面；第二个等号成立是根据假设 2.4 中的式 (2.5)；第三个等号成立是根据矩阵迹 (trace) 的定义。

接下来证明 (2)，由式 (A.4)，我们有：

$$\mathrm{tr}(\mathbf{M}_\mathbf{X}) = \mathrm{tr}(\mathbf{I}_n - \mathbf{P}_\mathbf{X}) = n - \mathrm{tr}(\mathbf{P}_\mathbf{X}),$$

同时，由式 (A.3)，

$$\mathrm{tr}(\mathbf{P}_\mathbf{X}) = \mathrm{tr}[\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'] = \mathrm{tr}[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}] = \mathrm{tr}(\mathbf{I}_K) = K.$$

因此，$\mathrm{tr}(\mathbf{M}_\mathbf{X}) = n-K$。 □

有了 σ^2 的估计，我们可以将 z_k 表达式中的 σ^2 替换为 s^2，将替换后的检验统计量称为 t 统计量，我们有如下定理：

定理 2.5. t 统计量的分布

考虑原假设 $\mathbb{H}_0$：$\beta_k = \beta_k^0$，此假设检验问题的 t 统计量被定义为：

$$t_k \equiv \frac{\hat{\beta}_{k,ols} - \beta_k^0}{\sqrt{s^2 \cdot \left((\mathbf{X}'\mathbf{X})^{-1}\right)_{kk}}}, \tag{2.23}$$

其中，$\sqrt{s^2 \cdot \left((\mathbf{X}'\mathbf{X})^{-1}\right)_{kk}}$ 称为标准误 (standard error)，通常记为 $SE(\hat{\beta}_{k,ols})$。令假设 2.1-2.5 成立，在原假设 $\mathbb{H}_0$ 下，t 统计量服从自由度为 $n-K$ 的 t 分布，记为 t_{n-K}。[a]

[a] 如果 $x \sim \mathcal{N}(0,1)$，$y \sim \chi^2(m)$ 并且 x 和 y 相互独立，则比值 $x/\sqrt{y/m}$ 服从自由度为 m 的 t 分布。 ♡

证明 t_k 可以表示为：

$$t_k = \frac{\hat{\beta}_{k,ols} - \beta_k^0}{\sqrt{\sigma^2 \cdot ((\mathbf{X}'\mathbf{X})^{-1})_{kk}}} \cdot \sqrt{\frac{\sigma^2}{s^2}} \equiv \frac{z_k}{\sqrt{s^2/\sigma^2}} = \frac{z_k}{\sqrt{\frac{\boldsymbol{e}'\boldsymbol{e}/(n-K)}{\sigma^2}}} \equiv \frac{z_k}{\sqrt{\frac{q}{n-K}}},$$

其中，$q = \boldsymbol{e}'\boldsymbol{e}/\sigma^2$。因为，$z_k$ 服从标准正态分布，我们只需证明：

(1) $q \mid \mathbf{X} \sim \chi^2(n-K)$；

(2) 以 $\mathbf{X}$ 为条件，z_k 和 q 相互独立。

首先，注意到：

$$q = \frac{\boldsymbol{e}'\boldsymbol{e}}{\sigma^2} = \frac{\boldsymbol{\varepsilon}'}{\sigma}\mathbf{M}_{\mathbf{X}}\frac{\boldsymbol{\varepsilon}}{\sigma}.$$

因为，$\mathbf{M}_{\mathbf{X}}$ 是对称幂等阵且 $\boldsymbol{\varepsilon}/\sigma \mid \mathbf{X} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_n)$，根据定理 B.12，我们有：$q \mid \mathbf{X} \sim \chi^2(\mathrm{rank}(\mathbf{M}_{\mathbf{X}}))$。又因为 $\mathbf{M}_{\mathbf{X}}$ 的秩（rank）等于它的迹（见 A.17 节的结论 12），所以，由定理 2.4 的证明，我们有：$\mathrm{rank}(\mathbf{M}) = n - K$。

接下来，我们证明 (2)。由式 (2.17) 和 (2.18)，$\hat{\boldsymbol{\beta}}_{ols}$ 和 $\boldsymbol{e}$ 都是 $\boldsymbol{\varepsilon}$ 的线性函数，因此，以 $\mathbf{X}$ 为条件，$\hat{\boldsymbol{\beta}}_{ols}$ 和 $\boldsymbol{e}$ 服从联合正态分布。再由定理 2.2 中 (d) 的结论，$\hat{\boldsymbol{\beta}}_{ols}$ 和 $\boldsymbol{e}$ 以 $\mathbf{X}$ 为条件彼此不相关，因此满足条件独立。注意到 z_k 是 $\hat{\boldsymbol{\beta}}_{ols}$ 的函数，q 是 $\boldsymbol{e}$ 的函数，所以，z_k 和 q 也满足条件独立性。

由 (1) 和 (2)，我们知道，以 $\mathbf{X}$ 为条件，t_k 服从自由度为 $n-K$ 的 t 分布，由于此分布与 $\mathbf{X}$ 无关，所以 t_k 的无条件分布也是 t_{n-K}。 □

基于 t 统计量的假设检验称为 t 检验，其具体步骤如下：

1. 根据原假设 $\mathbb{H}_0$ 中参数的具体取值 β_k^0 计算由式 (2.23) 定义的 t 统计量；
2. 通过统计软件计算自由度为 $n-K$ 的 t 分布的临界值 $t_{\alpha/2}(n-K)$（也可以通过查表的方式找到临界值），如图 2.4 所示，$t_{\alpha/2}(n-K)$ 满足 t 分布在此临界值右侧的阴影面积为 $\alpha/2$。在 R 语言中，我们可以通过命令 qt(1-α/2, $n-K$) 求得 $t_{\alpha/2}(n-K)$，例如，$\alpha = 5\%$，$n-K = 100$，我们可以用命令 qt(0.975, 100) 得到临界值为 1.984；
3. 如果第一步计算的 t_k 满足 $|t_k| < t_{\alpha/2}(n-K)$，则接受 $\mathbb{H}_0$；否则，拒绝 $\mathbb{H}_0$。

因为，在 $\mathbb{H}_0$ 成立的情况下，基于上述步骤的假设检验拒绝 $\mathbb{H}_0$ 的概率为 α，所以，α 是检验的显著性水平，英文文献中常称其为检验的 size，也就是犯第一类错误的概率。

我们还可以定义上述假设检验问题的置信水平为 $1-\alpha$ 的置信区间。根据定理 2.5，t 检验的不拒绝区域为：

$$\mathcal{D} = \left\{ (\mathbf{X}, \mathbf{y}) \ : \ \left| \frac{\hat{\beta}_{k,ols} - \beta_k^0}{SE(\hat{\beta}_{k,ols})} \right| < t_{\alpha/2}(n-K) \right\}.$$

通过求解上式中的不等式并将 β_k^0 用 β_k^* 代替，我们可知参数 β_k^* 的取值满足：

$$\hat{\beta}_{k,ols} - SE(\hat{\beta}_{k,ols}) \cdot t_{\alpha/2}(n-K) < \beta_k^* < \hat{\beta}_{k,ols} + SE(\hat{\beta}_{k,ols}) \cdot t_{\alpha/2}(n-K).$$

因此，我们接受 $\mathbb{H}_0$ 当且仅当原假设中的参数取值 β_k^0 落入区间：

$$[\hat{\beta}_{k,ols} - SE(\hat{\beta}_{k,ols}) \cdot t_{\alpha/2}(n-K), \quad \hat{\beta}_{k,ols} + SE(\hat{\beta}_{k,ols}) \cdot t_{\alpha/2}(n-K)],$$

这一区间被称为置信水平为 $1-\alpha$ 的置信区间，它覆盖真实参数的概率为 $1-\alpha$。

t 检验也可以通过计算 p 值判断是否接受 $\mathbb{H}_0$，具体步骤如下：

1. 根据原假设 $\mathbb{H}_0$ 中参数的具体取值 β_k^0 计算由式 (2.23) 定义的 t 统计量 t_k；

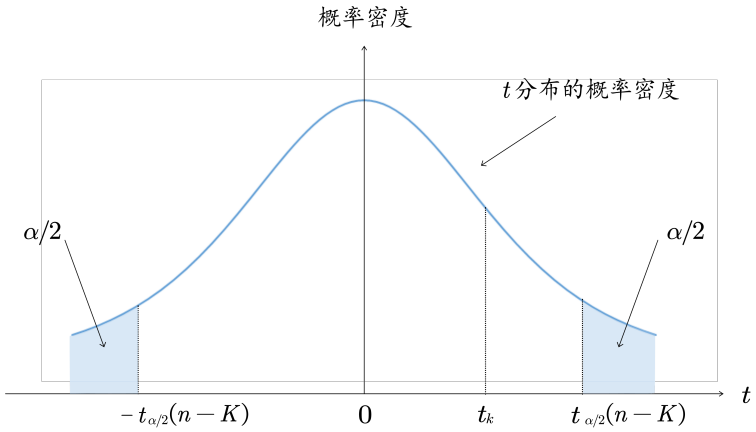

图 2.4: t 分布

2. 根据如下公式计算 p 值：

$$p = \mathbb{P}(t > |t_k|) \times 2, \tag{2.24}$$

其中，$\mathbb{P}(t > |t_k|)$ 是服从 t_{n-K} 分布的随机变量 t 取值大于 $|t_k|$ 的概率；

3. 如果 $p > \alpha$，则接受 $\mathbb{H}_0$；否则，拒绝 $\mathbb{H}_0$。

例 2.1 我们利用 French 个人主页[5]提供的 Fama-French 三因子的数据以及按照规模（market equity, ME）和账面市值比（BE/ME）的五分位点划分的 25 个资产组合中的小市值-低账面市值比组合的数据说明如何在 Stata 中实现单个回归系数的假设检验。样本时间范围是 1960-01-04 至 2020-03-31，样本量为 15164。读者可在本书网站下载数据预处理程序。本例中用到的主要 Stata 代码如下：

```
. use Fama-French_Data.dta, replace
. gen smalllobm_rf = smalllobm - rf
  //计算小市值-低账面市值比组合的超额收益率，rf为无风险收益率
. reg smalllobm_rf mktrf smb hml //smalllobm_rf对三因子回归
. gen t = _b[hml]/_se[hml] //计算HML因子系数显著性检验的t统计量
. gen p = 2*ttail(e(df_r),abs(t))
  //计算p值，e(df_r)是自由度，ttail命令用来计算t分布尾部的概率
. display p
.01291774
. test hml = 0 //检验HML的系数是否为0
 ( 1)  hml = 0
       F(  1, 15160) =   6.18
            Prob > F =   0.0129
. test hml = 1 //检验HML的系数是否为1
 ( 1)  hml = 1
```

[5] 网址为 http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/index.html。

```
14        F(  1, 15160) =14371.40
15             Prob > F =   0.0000
```

在上面程序的第 3 行，我们使用 reg 命令将小市值-低账面市值比组合的超额收益率对三因子进行回归，Stata 输出结果如下：

表 2.1: 例 2.1 回归结果

变量名	系数估计量	标准误	t 统计量	p 值
mktrf	1.086	0.004	250.824	0.000
smb	1.141	0.008	138.810	0.000
hml	-0.021	0.009	-2.486	0.013
_cons	-0.025	0.004	-6.036	0.000

其中，“t 统计量”一列对应的是系数显著性检验[6]的 t 值，“p 值”一列为各个 t 统计量对应的 p 值。程序的第 4 行利用式 (2.23) 计算 HML 因子系数显著性检验的 t 值；程序的第 5 行利用式 (2.24) 计算 HML 因子系数显著性检验的 p 值，其输出结果为 0.01291774，经四舍五入为 0.013，与表2.1中结果一致。

在 Stata 中，我们还可以利用 test 命令做线性假设检验。例如，程序的第 8 行检验 HML 因子的系数是否为 0，输出的 p 值与 reg 输出结果一致。但是，注意 test 命令输出的检验统计量是 F 统计量，它等于 t 统计量的平方（$-2.486^2 \approx 6.18$）。除了做显著性检验之外，我们还可以利用 test 命令检验单个回归系数是否等于某一常数，如程序的第 12 行，我们检验 HML 因子的系数是否等于 1。

一般线性假设检验

除了单个回归系数的假设检验问题，我们也可以考虑含有多个回归系数的线性假设检验问题。例如，考虑 Fama-French 三因子模型（见 Fama et al. (1992)）：

$$r_{p,t} = \alpha_p + \beta_{MKT,\,p} MKT_t + \beta_{SMB,\,p} SMB_t + \beta_{HML,\,p} HML_t + \varepsilon_{p,t},$$

其中，$r_{p,t}$ 是资产组合 p 在 t 时段的超额收益率；MKT_t、SMB_t 和 HML_t 分别表示市场因子、规模因子和价值因子[7]。三因子对资产组合 p 的超额收益率是否有解释能力可以通过如下假设检验进行分析：

$$\mathbb{H}_0 \ : \ \beta_{MKT,\,p} = \beta_{SMB,\,p} = \beta_{HML,\,p} = 0.$$

备择假设为：

$$\mathbb{H}_1 \ : \ \beta_{MKT,\,p},\ \beta_{SMB,\,p} \text{ 和 } \beta_{HML,\,p} \text{ 至少有一个不为 } 0。$$

[6] 系数显著性检验的原假设是解释变量的系数取值为 0。

[7] 关于三因子的具体含义和构造方式读者可参考 Bali et al. (2016)。

由于原假设是除常数项以外的所有解释变量的系数为 0，因此，此类假设检验问题被称为联合显著性检验。我们可以把上述假设检验表示为一般形式，为此，定义如下矩阵：

$$\mathbf{R} = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

以及向量 $\mathbf{r} = (0, 0, 0)'$，则有：

$$\mathbb{H}_0 \ : \ \mathbf{R}\boldsymbol{\beta} = \mathbf{r},$$

其中，$\boldsymbol{\beta} = (\alpha_p, \beta_{MKT,\,p}, \beta_{SMB,\,p}, \beta_{HML,\,p})'$。

事实上，所有关于多元线性回归系数的线性假设检验的原假设都可以表示为上述一般形式。例如，如果将 $\mathbf{R}$ 取为第 k 个位置上元素为 1 其余位置上元素为 0 的 K 维行向量以及取 $\mathbf{r} = \beta_k^0$，则单个回归系数的假设检验的原假设可以表示为上述一般形式。

对于一般形式的线性假设检验，我们有如下结论：

定理 2.6. 线性假设检验的 F 统计量

考虑原假设 $\mathbb{H}_0 : \mathbf{R}\boldsymbol{\beta} = \mathbf{r}$，其中，$\mathbf{r}$ 是 $J \times 1$ 维列向量；$\mathbf{R}$ 是 $J \times K$ 维矩阵且满足 $\mathrm{rank}(\mathbf{R}) = J$，此假设检验问题的 F 统计量被定义为：

$$\begin{aligned} F &\equiv \frac{(\mathbf{R}\hat{\boldsymbol{\beta}}_{ols} - \mathbf{r})' \left[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'\right]^{-1} (\mathbf{R}\hat{\boldsymbol{\beta}}_{ols} - \mathbf{r})/J}{s^2} \\ &= (\mathbf{R}\hat{\boldsymbol{\beta}}_{ols} - \mathbf{r})' \left[\mathbf{R}\widehat{\mathrm{Var}(\hat{\boldsymbol{\beta}}_{ols}|\mathbf{X})}\mathbf{R}'\right]^{-1} (\mathbf{R}\hat{\boldsymbol{\beta}}_{ols} - \mathbf{r})/J. \end{aligned} \tag{2.25}$$

令假设 2.1-2.5 成立，在原假设 $\mathbb{H}_0$ 下，F 统计量服从自由度为 J 和 $n-K$ 的 F 分布，记为 $F(J, n-K)$。[a]

[a] 若 $x_1 \sim \chi^2(n_1)$，$x_2 \sim \chi^2(n_2)$ 并且 x_1 和 x_2 相互独立，则比值 $(x_1/n_1)/(x_2/n_2)$ 服从自由度为 n_1 和 n_2 的 F 分布。

♡

证明 因为，$s^2 = \boldsymbol{e}'\boldsymbol{e}/(n-K)$，则有：

$$F = \frac{w/J}{q/(n-K)},$$

其中，

$$w \equiv (\mathbf{R}\hat{\boldsymbol{\beta}}_{ols} - \mathbf{r})' \left[\sigma^2 \cdot \mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'\right]^{-1} (\mathbf{R}\hat{\boldsymbol{\beta}}_{ols} - \mathbf{r}), \quad q \equiv \frac{\boldsymbol{e}'\boldsymbol{e}}{\sigma^2}.$$

我们只需证明：

(1) $w \mid \mathbf{X} \sim \chi^2(J)$；

(2) $q \mid \mathbf{X} \sim \chi^2(n-K)$；

(3) 以 $\mathbf{X}$ 为条件，w 和 q 相互独立。

首先证明 (1)，令 $\mathbf{v} \equiv \mathbf{R}\hat{\boldsymbol{\beta}}_{ols} - \mathbf{r}$，若 $\mathbb{H}_0$ 成立，则有 $\mathbf{v} = \mathbf{R}(\hat{\boldsymbol{\beta}}_{ols} - \boldsymbol{\beta})$。由式 (2.22)，

$$\mathbf{v} \mid \mathbf{X} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \cdot \mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}').$$

因此，由定理 B.13，我们有 $w \mid \mathbf{X} \sim \chi^2(J)$。

结论 (2) 在定理 2.5 已经证明，我们接下来证明 (3)：注意到 w 是 $\hat{\boldsymbol{\beta}}_{ols}$ 的函数，q 是 $\boldsymbol{e}$ 的函数，与定理 2.5 的证明类似，容易验证 w 和 q 条件独立。

由 (1)-(3)，我们知道，以 $\mathbf{X}$ 为条件，F 统计量服从 $F(J, n-K)$ 分布，由于此分布与 $\mathbf{X}$ 无关，所以 F 统计量的无条件分布也是 $F(J, n-K)$。 □

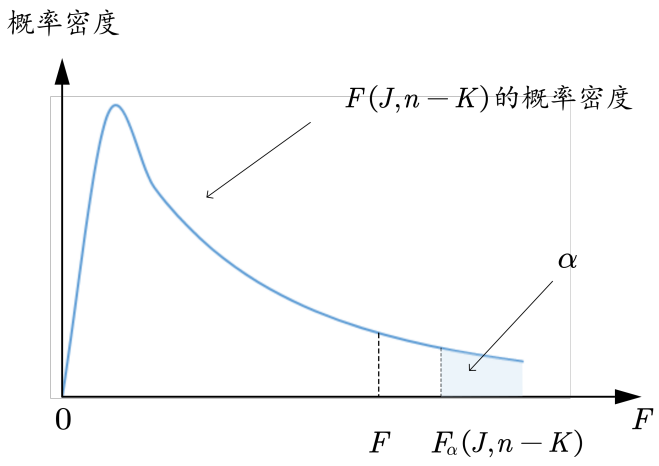

图 2.5: F 分布

从式 (2.25) 可以看出，F 统计量实际上计算了 $\mathbf{R}\hat{\boldsymbol{\beta}}_{ols}$ 和原假设下的真实值 $\mathbf{R}\boldsymbol{\beta}$ 对应元素之间距离的加权平方和，即，$\mathbf{R}\hat{\boldsymbol{\beta}}_{ols}$ 与 $\mathbf{R}\boldsymbol{\beta}$ 的马氏距离（Mahalanobis distance）的平方，若基于样本观测值计算出的此距离过大，则表明原假设可能不成立。因为距离是非负的，所以，在基于 F 统计量做假设检验时，临界值的选取只依赖于 F 分布的右尾部，如图 2.5。基于 F 统计量的假设检验称为 F 检验，其具体步骤如下：

1. 根据式 (2.25) 计算 F 统计量；
2. 通过统计软件计算 $F(J, n-K)$ 的临界值 $F_\alpha(J, n-K)$（也可以通过查表的方式找到临界值），如图 2.5 所示，$F_\alpha(J, n-K)$ 满足 F 分布在此临界值右侧的阴影面积为 α。在 R 语言中，我们可以通过命令 qf(1-α, J, $n-K$) 求得 $F_\alpha(J, n-K)$，例如，$\alpha = 5\%$，$J = 3$，$n-K = 30$，我们可以用命令 qf(0.95, 3, 30) 得到临界值为 2.922；
3. 如果第一步计算的 F 统计量满足 $F < F_\alpha(J, n-K)$，则接受 $\mathbb{H}_0$；否则，拒绝 $\mathbb{H}_0$。

我们也可以根据 p 值判断接受还是拒绝 $\mathbb{H}_0$，具体步骤如下：

1. 根据式 (2.25) 计算 F 统计量；
2. 计算分布 $F(J, n-K)$ 在 F 统计量右侧区域的面积 p；
3. 如果 $p > \alpha$，则接受 $\mathbb{H}_0$；否则，拒绝 $\mathbb{H}_0$。

基于残差平方和的 F 统计量表达式

F 统计量的表达式 (2.25) 还有另外一种等价的表达形式，涉及到有约束和无约束回归的残差平方和。

定理 2.7. F 统计量的等价形式

F 统计量可以表示为：

$$F = \frac{(RSS_R - RSS_U)/J}{RSS_U/(n-K)},$$

其中，RSS_U 是无约束回归的残差平方和，即，$\boldsymbol{e}'\boldsymbol{e}$；$RSS_R$ 是有约束回归的残差平方和，由求解如下最优化问题得到：

$$\min_{\boldsymbol{b}} RSS(\boldsymbol{b}) \quad \text{s.t.} \quad \mathbf{R}\boldsymbol{b} = \mathbf{r}.$$

♡

证明 令 $\tilde{\boldsymbol{\beta}}$ 是在 $\mathbb{H}_0$ 约束下的 OLS 估计，即，

$$\tilde{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{b}} (\mathbf{y} - \mathbf{X}\boldsymbol{b})'(\mathbf{y} - \mathbf{X}\boldsymbol{b}) \quad \text{s.t.} \quad \mathbf{R}\boldsymbol{b} = \mathbf{r}.$$

构造上述有约束最优化问题的拉格朗日函数：

$$L(\boldsymbol{b}, \boldsymbol{\lambda}) = (\mathbf{y} - \mathbf{X}\boldsymbol{b})'(\mathbf{y} - \mathbf{X}\boldsymbol{b}) + 2\boldsymbol{\lambda}'(\mathbf{r} - \mathbf{R}\boldsymbol{b}),$$

其中，$\boldsymbol{\lambda}$ 是 $J \times 1$ 维的拉格朗日乘数向量。由最优化的一阶条件 (C.6)，我们有：

$$\frac{\partial L(\boldsymbol{b}, \boldsymbol{\lambda})}{\partial \boldsymbol{b}}\Big|_{\boldsymbol{b}=\tilde{\boldsymbol{\beta}},\, \boldsymbol{\lambda}=\tilde{\boldsymbol{\lambda}}} = -2\mathbf{X}'(\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}}) - 2\mathbf{R}'\tilde{\boldsymbol{\lambda}} = \mathbf{0}, \tag{2.26}$$

$$\frac{\partial L(\boldsymbol{b}, \boldsymbol{\lambda})}{\partial \boldsymbol{\lambda}}\Big|_{\boldsymbol{b}=\tilde{\boldsymbol{\beta}}} = 2(\mathbf{r} - \mathbf{R}\tilde{\boldsymbol{\beta}}) = \mathbf{0}. \tag{2.27}$$

由式 (2.26)，我们有：

$$\begin{aligned}
-\mathbf{X}'(\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}}) &= -\mathbf{X}'(\mathbf{X}\hat{\boldsymbol{\beta}}_{ols} + \boldsymbol{e} - \mathbf{X}\tilde{\boldsymbol{\beta}}) && (\text{因为 } \mathbf{y} = \mathbf{X}\hat{\boldsymbol{\beta}}_{ols} + \boldsymbol{e}) \\
&= -\mathbf{X}'\mathbf{X}(\hat{\boldsymbol{\beta}}_{ols} - \tilde{\boldsymbol{\beta}}) && (\text{因为 } \mathbf{X}'\boldsymbol{e} = \mathbf{0}) \\
&= \mathbf{R}'\tilde{\boldsymbol{\lambda}}.
\end{aligned}$$

因此，

$$-(\hat{\boldsymbol{\beta}}_{ols} - \tilde{\boldsymbol{\beta}}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'\tilde{\boldsymbol{\lambda}}, \tag{2.28}$$

两边同时左乘 $\mathbf{R}$ 得到：$-\mathbf{R}(\hat{\boldsymbol{\beta}}_{ols} - \tilde{\boldsymbol{\beta}}) = \mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'\tilde{\boldsymbol{\lambda}}$，再根据 $\tilde{\boldsymbol{\beta}}$ 满足式 (2.27)，我们有：

$$\tilde{\boldsymbol{\lambda}} = -[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{R}\hat{\boldsymbol{\beta}}_{ols} - \mathbf{r}).$$

将上式代入 (2.28)，则有：

$$\hat{\boldsymbol{\beta}}_{ols} - \tilde{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{R}\hat{\boldsymbol{\beta}}_{ols} - \mathbf{r}).$$

另一方面，有约束回归的残差可以表示为：

$$\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_{ols} + \mathbf{X}(\hat{\boldsymbol{\beta}}_{ols} - \tilde{\boldsymbol{\beta}}) = \boldsymbol{e} + \mathbf{X}(\hat{\boldsymbol{\beta}}_{ols} - \tilde{\boldsymbol{\beta}}).$$

于是，有约束回归的残差平方和为：

$$\begin{aligned}
RSS_R &= (\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}})'(\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}}) = [\boldsymbol{e} + \mathbf{X}(\hat{\boldsymbol{\beta}}_{ols} - \tilde{\boldsymbol{\beta}})]'[\boldsymbol{e} + \mathbf{X}(\hat{\boldsymbol{\beta}}_{ols} - \tilde{\boldsymbol{\beta}})] \\
&= \boldsymbol{e}'\boldsymbol{e} + 2(\hat{\boldsymbol{\beta}}_{ols} - \tilde{\boldsymbol{\beta}})'\mathbf{X}'\boldsymbol{e} + (\hat{\boldsymbol{\beta}}_{ols} - \tilde{\boldsymbol{\beta}})'\mathbf{X}'\mathbf{X}(\hat{\boldsymbol{\beta}}_{ols} - \tilde{\boldsymbol{\beta}}) \\
&= \boldsymbol{e}'\boldsymbol{e} + (\hat{\boldsymbol{\beta}}_{ols} - \tilde{\boldsymbol{\beta}})'\mathbf{X}'\mathbf{X}(\hat{\boldsymbol{\beta}}_{ols} - \tilde{\boldsymbol{\beta}}) && (\mathbf{X}'\boldsymbol{e} = \mathbf{0}) \\
&= \boldsymbol{e}'\boldsymbol{e} + (\mathbf{R}\hat{\boldsymbol{\beta}}_{ols} - \mathbf{r})'[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{R}\hat{\boldsymbol{\beta}}_{ols} - \mathbf{r}).
\end{aligned}$$

由上式，我们有：

$$(\mathbf{R}\hat{\boldsymbol{\beta}}_{ols} - \mathbf{r})'[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{R}\hat{\boldsymbol{\beta}}_{ols} - \mathbf{r}) = RSS_R - RSS_U.$$

因此，

$$
\begin{aligned}
F &= \frac{(\mathbf{R}\hat{\boldsymbol{\beta}}_{ols}-\mathbf{r})'\left[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'\right]^{-1}(\mathbf{R}\hat{\boldsymbol{\beta}}_{ols}-\mathbf{r})/J}{s^2} \\
&= \frac{(RSS_R-RSS_U)/J}{RSS_U/(n-K)}.
\end{aligned}
$$

□

例 2.2 考虑如下线性回归模型：

$$y_i=\beta_1+\beta_2x_{i2}+\cdots+\beta_Kx_{iK}+\varepsilon_i,$$

和原假设 $\mathbb{H}_0:\beta_j=0$，$2\leqslant j\leqslant K$，即检验解释变量的联合显著性，此时 $J=K-1$。由式 (2.11)，有约束回归的残差平方和为：

$$RSS_R=\sum_{i=1}^{n}(y_i-\bar{y})^2.$$

于是，由 R^2 的定义，我们有：

$$R^2=1-\frac{\boldsymbol{e}'\boldsymbol{e}}{\sum_{i=1}^{n}(y_i-\bar{y})^2}=1-\frac{\boldsymbol{e}'\boldsymbol{e}}{RSS_R}.$$

于是，

$$
\begin{aligned}
F &= \frac{(RSS_R-\boldsymbol{e}'\boldsymbol{e})/(K-1)}{\boldsymbol{e}'\boldsymbol{e}/(n-K)}=\frac{\left(1-\dfrac{\boldsymbol{e}'\boldsymbol{e}}{RSS_R}\right)/(K-1)}{\dfrac{\boldsymbol{e}'\boldsymbol{e}}{RSS_R}/(n-K)} \\
&= \frac{R^2/(K-1)}{(1-R^2)/(n-K)}. \qquad (2.29)
\end{aligned}
$$

例 2.3 我们对例2.1中考虑的 Fama-French 三因子模型进行解释变量的联合显著性检验，所使用的数据集与例2.1完全相同。此时，$J=3$，$n-K=15164-4=15160$。在 Stata 中，reg 命令的输出结果包含了解释变量的联合显著性检验的 F 统计量取值和对应的 p 值。例2.1的第 3 行程序的输出结果如图 2.6 所示。图片右上方椭圆内汇报的便是解释变量联合显著性检验的检验结果，F 统计量服从 $F(3,15160)$ 分布，其取值等于 26081.55，检验的 p 值近似为 0，因此，我们可以拒绝原假设。我们可以通过式 (2.29) 计算 F 统计量。首先，通过 Stata 命令 ereturn list 获取 e-class 的返回值，其中，R^2 的返回值为：e(r2) = .8376954571692361。将此数值代入到式 (2.29) 中便可得到 F 统计量的取值为 26081.55。

F 检验的置信域

因为 F 统计量本质上度量了 $\mathbf{R}\hat{\boldsymbol{\beta}}_{ols}$ 和 $\mathbf{R}\boldsymbol{\beta}$ 之间的马氏距离，所以，F 检验的置信域与马氏距离及其等高线（contour line）有关。本章网上附录对统计（马氏）距离做了较为详细地介绍，建议不熟悉相关知识的读者在阅读本小节内容前先阅读该附录。

由定理 2.6，F 检验的不拒绝区域为：

$$\mathcal{D}=\left\{(\mathbf{X},\mathbf{y}):(\mathbf{R}\hat{\boldsymbol{\beta}}_{ols}-\mathbf{r})'\left[\mathbf{R}\widehat{\mathrm{Var}(\hat{\boldsymbol{\beta}}_{ols}|\mathbf{X})}\mathbf{R}'\right]^{-1}(\mathbf{R}\hat{\boldsymbol{\beta}}_{ols}-\mathbf{r})<J\cdot F_\alpha(J,n-K)\right\}.$$

根据统计距离附录的介绍，我们知道：若原假设中参数取值 $\boldsymbol{\beta}$ 满足 $\mathbf{r}=\mathbf{R}\boldsymbol{\beta}$ 落入以 $\mathbf{R}\hat{\boldsymbol{\beta}}_{ols}$

```
. reg smalllobm_rf mktrf smb hml

      Source |       SS           df       MS      Number of obs   =    15,164
-------------+----------------------------------   F(3, 15160)     =  26081.55
       Model |  20951.8333         3  6983.94444   Prob > F        =    0.0000
    Residual |  4059.44392    15,160  .267773345   R-squared       =    0.8377
-------------+----------------------------------   Adj R-squared   =    0.8377
       Total |  25011.2772    15,163  1.64949398   Root MSE        =    .51747

------------------------------------------------------------------------------
smalllobm_rf | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
       mktrf |   1.085563    .004328   250.82   0.000      1.07708    1.094046
         smb |   1.141249   .0082217   138.81   0.000     1.125134    1.157364
         hml |  -.0211792   .0085183    -2.49   0.013    -.0378761   -.0044824
       _cons |  -.0253872    .004206    -6.04   0.000    -.0336314    -.017143
------------------------------------------------------------------------------
```

图 2.6: 三因子联合显著性检验

为中心的超椭球

$$\left\{\mathbf{r}\in\mathbb{R}^J\ :\ (\mathbf{R}\hat{\boldsymbol{\beta}}_{ols}-\mathbf{r})'\left[\mathbf{R}\widehat{\mathrm{Var}(\hat{\boldsymbol{\beta}}_{ols}|\mathbf{X})}\mathbf{R}'\right]^{-1}(\mathbf{R}\hat{\boldsymbol{\beta}}_{ols}-\mathbf{r})=J\cdot F_\alpha(J,n-K)\right\}$$

内部，则原假设不能被拒绝。据此，我们可以得到 $\mathbb{H}_0$ 中涉及参数的置信域。

例如：假设 $K=2$，考虑如下原假设：

$$\mathbb{H}_0\ :\ \beta_1=\beta_1^0,\quad \beta_2=\beta_2^0.$$

此时，我们有 $\mathbf{R}=\mathbf{I}_2$ 和 $\mathbf{r}=(\beta_1^0,\beta_2^0)'$。$F$ 检验的置信水平为 $1-\alpha$ 的置信域为：

$$\left\{\begin{pmatrix}\beta_1\\ \beta_2\end{pmatrix}\ :\ \begin{pmatrix}\hat{\beta}_{1,ols}-\beta_1\\ \hat{\beta}_{2,ols}-\beta_2\end{pmatrix}'\left[\widehat{\mathrm{Var}(\hat{\boldsymbol{\beta}}_{ols}|\mathbf{X})}\right]^{-1}\begin{pmatrix}\hat{\beta}_{1,ols}-\beta_1\\ \hat{\beta}_{2,ols}-\beta_2\end{pmatrix}<2F_\alpha(2,n-K)\right\}.$$

如图 2.7，上述置信域是以 $(\hat{\beta}_{1,ols},\hat{\beta}_{2,ols})$ 为中心的椭圆内部区域。若图中 B 点为 (β_1^0,β_2^0)，则接受 $\mathbb{H}_0$；若图中 A 点为 (β_1^0,β_2^0)，则拒绝 $\mathbb{H}_0$。另一方面，一个值得思考的问题是此例中的原假设能否通过两个单系数的 t 统计量进行检验：若两个 t 检验都接受各自的原假设，则接受 $\mathbb{H}_0$；否则，拒绝 $\mathbb{H}_0$。若两个 t 检验的显著性水平都取为 α，则通过两个单系数的 t 统计量检验 $\mathbb{H}_0$ 的检验置信域为图 2.7 中两个参数的置信水平为 $1-\alpha$ 的置信区间围成的矩形区域。若 $\mathbb{H}_0$ 中参数取值对应图 2.7 中的 A 点，则 F 检验拒绝 $\mathbb{H}_0$，双系数 t 检验接受 $\mathbb{H}_0$；若 $\mathbb{H}_0$ 中参数取值对应图 2.7 中的 B 点，则 F 检验接受 $\mathbb{H}_0$，双系数 t 检验拒绝 $\mathbb{H}_0$。在实证分析时，若 $J\geqslant 2$，应该使用 F 检验，其原因主要体现在以下两点：首先，若单个系数的 t 检验的显著性水平为 α，则多系数联合 t 检验的显著性水平不等于 α；其次，由统计距离附录的分析可知，在构造 F 检验的置信域时，我们考虑了各系数估计量的相关性，而多系数联合 t 检验忽略了此相关性。因此，即使可以通过调整单个系数的 t 检验的显著性水平使得总体显著性水平控制在 α，实际中我们也更偏好使用 F 统计量检验 $J\geqslant 2$ 的原假设。

事实上，t 统计量也可以看成是 F 统计量的特例。考虑 t 检验的原假设 $\mathbb{H}_0\ :\ \beta_k=\beta_k^0$，我们可以取 $\underset{(1\times K)}{\mathbf{R}}=[\underbrace{0,\ldots,0}_{k-1\text{个}},1,0,\ldots,0]$ 和 $\mathbf{r}=\beta_k^0$，则 $\mathbf{R}\boldsymbol{\beta}=\mathbf{r}$ 等价于 $\beta_k=\beta_k^0$。在 $\mathbb{H}_0$ 成

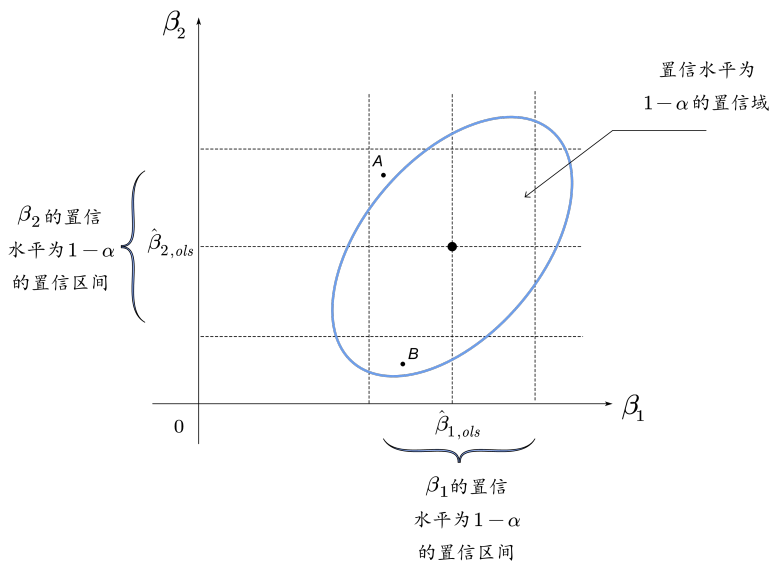

图 2.7: t 检验与 F 检验

立的情况下，F 统计量服从 $F(1, n-K)$ 分布；t 统计量服从 t_{n-k} 分布。因为 $F(1, n-K)$ 分布是 t_{n-k} 分布的平方，所以，t 检验和 F 检验的结果一致。

2.4 多重共线性

假设 2.3 要求矩阵 $\mathbf{X}$ 列满秩，否则 $\mathbf{X}'\mathbf{X}$ 不可逆，最小二乘估计量无法计算。若 $\mathbf{X}$ 的各列之间存在高度的（不完全的）线性相关关系，最小二乘估计量仍然可以计算，但估计量的方差将增大从而影响估计量的准确性。出现这种现象的原因是当解释变量之间高度线性相关时矩阵 $\mathbf{X}'\mathbf{X}$ 接近奇异，利用计算机无法准确求逆，当数据矩阵 $\mathbf{X}$ 的元素发生微小变化时，系数估计量的数值变化很大，甚至出现符号相反的情况。

我们通过数值模拟的方式来更清楚地说明多重共线性问题。首先，假设数据矩阵 $\mathbf{X}$ 和 $\mathbf{y}$ 为：

$$\mathbf{X} = \begin{bmatrix} 1 & 0.98 \\ 1 & 0.99 \\ 1 & 0.99 \end{bmatrix}, \quad \mathbf{y} = \begin{bmatrix} 3.5 \\ 2.6 \\ 4.9 \end{bmatrix}.$$

由最小二乘估计量公式 (2.8)，我们有：

$$\hat{\boldsymbol{\beta}}_{ols} = (-21,\ 25)'.$$

现假设

$$\mathbf{X} = \begin{bmatrix} 1 & 0.98 \\ 1 & 0.99 \\ 1 & 0.98 \end{bmatrix},$$

则最小二乘估计量为：

$$\hat{\boldsymbol{\beta}}_{ols} = (161,\ -160)'.$$

通过数值模拟，我们发现：在数据矩阵两列高度线性相关的情况下，第三个样本的解释变量观测值变化 0.01 导致最小二乘估计量的绝对值发生很大的变化，而且符号相反。因此，多重共线性是一种样本现象，不是线性回归模型本身的问题，可以通过调整样本减轻多重共线性的影响。

由式 (2.20)，我们知道 β_j 的估计量的方差与 R_j^2 有关。R_j^2 衡量的是回归中其他解释变量可以解释的 x_j 变动占 x_j 总变动的比例。只要假设 2.3 成立，R_j^2 不会等于 1。但是，x_j 与其他解释变量高度线性相关会导致 R_j^2 非常接近于 1。由式 (2.20)，随着 R_j^2 趋近于 1，$\hat{\beta}_{j,ols}$ 的方差将趋于无穷，从而导致多重共线性问题。由 t 统计量公式 (2.23)，若估计量方差很大，则其标准误也将很大，导致 t 统计量可能不显著。所以，如果系数的联合显著性 F 检验显著，但是单个参数 t 检验都不显著，则很可能出现了多重共线性问题。除此之外，多重共线性还可能导致如下现象：

- 最小二乘估计量的数值或符号与经济金融理论不符；
- 如果增加某一解释变量，R^2 无明显变化，但是 t 统计量绝对值明显减小；
- 如果删除某一解释变量，R^2 无明显变化，但是 t 统计量绝对值明显变大；
- 样本的增减导致最小二乘估计量较大的变化。

但是，多重共线性问题是否会导致一个过大的系数估计量方差还取决于 σ^2 和 TSS_j。而且，t 统计量是否受影响还取决于系数估计量与其标准误的比值。一般情况下，我们总是可以通过收集更多的数据来增加 TSS_j，从而减小估计量的方差。在样本固定的情况下，x_j 与其他解释变量的相关性越低越有助于准确估计 β_j，最理想的情况就是 x_j 与其他解释变量都不相关。但是多元线性回归中加入其他解释变量的原因正是因为 x_j 与其他解释变量存在相关性，若遗漏这些变量会导致假设 2.2 不成立。所以，即使解释变量间出现了高度的相关性，我们也不能随意的丢弃引起这种相关性的解释变量，除非我们在变量选择方面出现了问题。例如，我们想研究家庭各项教育投入对子女成绩的影响，解释变量包括补课费用、教辅资料费用、文具用品费用等。显然，这些解释变量之间具有高度的相关性，富裕家庭在每一项上都投入很多，贫困家庭在每一项上都有所节制。所以，我们很难准确估计这些投入项目对子女成绩的偏效应，这便是变量选择上出了问题。

此外，多元线性回归中的各个解释变量的地位通常是不同的。一般地，我们只关心某一核心解释变量对被解释变量的影响，其他解释变量是为了保证假设 2.2 成立而引入的控制变量。控制变量之间的高度线性相关性可能不会影响核心解释变量系数估计量的方差。例如，考虑如下回归：

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \varepsilon_i,$$

其中，控制变量 x_{2i} 和 x_{3i} 高度相关。虽然，$\mathrm{Var}(\hat{\beta}_{2,ols}|\mathbf{X})$ 和 $\mathrm{Var}(\hat{\beta}_{3,ols}|\mathbf{X})$ 可能很大，但是，只要 x_{1i} 与 (x_{2i}, x_{3i}) 的相关程度不高，$\mathrm{Var}(\hat{\beta}_{1,ols}|\mathbf{X})$ 不会受到 x_{2i} 和 x_{3i} 高度相关的影响。此时，若 β_1 是感兴趣的参数，我们大可不必担心 x_{2i} 和 x_{3i} 是高度相关的。例如，在研究金融开放程度与地方经济增长关系的回归分析中，我们可能会加入收入和房价等控制变量，虽然这些控制变量存在高度的正相关关系，但是只要金融开放程度与收入和房价的相关性不强，我们依然可以准确估计感兴趣的参数。

常用的识别多重共线性的方法还有成对（pairwise）样本相关系数表和方差膨胀因子

(variance inflation factor, VIF)。VIF 的定义为

$$\mathrm{VIF}_j = \frac{1}{1-R_j^2}, \quad j = 2, \ldots, K.$$

在 Stata 中，成对样本相关系数表可以使用命令 pwcorr 得到，VIF 可以在回归之后使用命令 estat vif 得到。通常，若变量的 VIF 大于 10，则认为该变量引起了多重共线性问题。然而，通过前文的分析，我们应该谨慎使用这两种方法进行变量的筛选。例如，发现两个控制变量的样本相关系数超过 0.9 也可能不影响我们关心的核心参数估计；若在核心解释变量和被解释变量的因果关系推断中我们必须控制某一变量，即使该变量的 VIF 大于 10，我们也不能随意丢弃它。所以，无论是成对样本相关系数表还是 VIF 在确定总体回归函数方面的作用都是有限的。若使用多元回归模型的目的是进行因果推断，那么模型中的变量选择的最终目标是保证核心参数的因果效应解释，而不仅仅是一个统计学问题。

在实证分析中，若我们关心的核心系数的估计显著且其符号与预期相符，我们可以不考虑多重共线性问题；若不关心模型的具体系数，而是使用多元线性回归模型进行对被解释变量的预测，我们也可以不考虑多重共线性问题；若核心解释变量的系数估计不显著的原因是多重共线性，我们可以通过增大样本量减轻多重共线性的影响；如果增大样本量后核心解释变量的系数估计依然不显著，我们需要考虑回归模型的设定是否合理；在变量筛选过程中，成对样本相关系数表和方差膨胀因子等统计指标只能起到辅助作用，不应作为起决定性作用的筛选标准。

2.5 广义最小二乘

假设 2.4 要求误差 ε 的条件方差 $\mathbb{E}(\varepsilon\varepsilon'|\mathbf{X})$ 满足球形扰动假设。若没有假设 2.4 的限制，$\mathbb{E}(\varepsilon\varepsilon'|\mathbf{X})$ 通常是 $\mathbf{X}$ 的非线性函数。若没有条件同方差假设，则 $\mathbb{E}(\varepsilon\varepsilon'|\mathbf{X})$ 的对角线元素可以不相等；若误差项之间存在相关性，则 $\mathbb{E}(\varepsilon\varepsilon'|\mathbf{X})$ 的非对角线元素可以不为 0。本节中，我们放松球形扰动假设，讨论多元线性回归的广义最小二乘（Generalized Least Squares, GLS）估计。

我们先考虑一种理想的情况：假设 $n \times n$ 维矩阵 $\mathbf{V}(\mathbf{X})$ 是非奇异的对称正定矩阵，并且 $\mathbf{V}$ 中的元素是已知的[8]。令

$$\mathbb{E}(\varepsilon\varepsilon'|\mathbf{X}) = \sigma^2\mathbf{V}(\mathbf{X}), \tag{2.30}$$

其中，σ^2 是未知的。用式 (2.30) 替换假设 2.4 得到的模型称为广义回归模型。放松假设 2.4 会使得最小二乘估计量的条件方差表达式发生变化。具体地，我们有：

$$\begin{aligned}
\mathrm{Var}(\hat{\boldsymbol{\beta}}_{ols}|\mathbf{X}) &= \mathrm{Var}(\hat{\boldsymbol{\beta}}_{ols} - \boldsymbol{\beta}|\mathbf{X}) && \text{(减去常数 } \boldsymbol{\beta} \text{ 不改变方差)} \\
&= \mathrm{Var}(\mathbf{A}\varepsilon|\mathbf{X}) && \text{(由式 (2.17)，其中 } \mathbf{A} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\text{)} \\
&= \mathbf{A}\mathbb{E}(\varepsilon\varepsilon'|\mathbf{X})\mathbf{A}' \\
&= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}(\mathbf{X})\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}. && \text{(将式 (2.30) 代入)}
\end{aligned}$$

[8] 为了书写方便，我们也将 $\mathbf{V}(\mathbf{X})$ 简单地记为 $\mathbf{V}$。

因此，Gauss-Markov 定理不再成立。此外，由于最小二乘估计量的条件方差表达式发生变化，t 统计量不服从 t 分布，F 统计量不服从 F 分布。然而，最小二乘估计量依然是无偏估计量，这是因为无偏性质的推导没有使用假设 2.4。

接下来，在 $\mathbf{V}$ 已知的情况下，我们将给出比最小二乘估计量更有效的估计量，即，广义最小二乘估计量。因为 $\mathbf{V}$ 是对称正定矩阵，由对称矩阵的谱分解定理，我们有：

$$\mathbf{V} = \mathbf{C}\boldsymbol{\Lambda}\mathbf{C}',$$

其中，$\mathbf{C}$ 是正交矩阵，由矩阵 $\mathbf{V}$ 的单位正交特征向量构成；$\boldsymbol{\Lambda}$ 是对角阵，对角线上元素是 $\mathbf{C}$ 中特征向量对应的特征值。因为，$\mathbf{V}$ 是正定阵，其特征值都为正数，因此，我们有：

$$\mathbf{V}^{-1} = \mathbf{P}'\mathbf{P},$$

其中，$\mathbf{P} \equiv \boldsymbol{\Lambda}^{-1/2}\mathbf{C}'$。现考虑用 $\mathbf{P}$ 对 $(\mathbf{y}, \mathbf{X}, \boldsymbol{\varepsilon})$ 做线性变换，得到如下新回归模型中的变量：

$$\tilde{\mathbf{y}} \equiv \mathbf{P}\mathbf{y}, \quad \tilde{\mathbf{X}} \equiv \mathbf{P}\mathbf{X}, \quad \tilde{\boldsymbol{\varepsilon}} \equiv \mathbf{P}\boldsymbol{\varepsilon}.$$

对于 $(\tilde{\mathbf{y}}, \tilde{\mathbf{X}}, \tilde{\boldsymbol{\varepsilon}})$，我们有：

- 线性假设 2.1 成立：

$$\tilde{\mathbf{y}} = \tilde{\mathbf{X}}\boldsymbol{\beta} + \tilde{\boldsymbol{\varepsilon}}. \tag{2.31}$$

- 严格外生性假设 2.2 成立：

$$\mathbb{E}(\tilde{\boldsymbol{\varepsilon}} \mid \tilde{\mathbf{X}}) = \mathbb{E}(\tilde{\boldsymbol{\varepsilon}} \mid \mathbf{X}) = \mathbf{P}\mathbb{E}(\boldsymbol{\varepsilon} \mid \mathbf{X}) = \mathbf{0}.$$

- 因为 $\mathbf{P}$ 非奇异，无完全共线性假设 2.3 成立。
- 球形扰动假设 2.4 成立：

$$\begin{aligned}\mathbb{E}(\tilde{\boldsymbol{\varepsilon}}\tilde{\boldsymbol{\varepsilon}}' \mid \tilde{\mathbf{X}}) &= \mathbb{E}(\tilde{\boldsymbol{\varepsilon}}\tilde{\boldsymbol{\varepsilon}}' \mid \mathbf{X}) = \mathbf{P}\mathbb{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}' \mid \mathbf{X})\mathbf{P}' \\ &= \sigma^2\mathbf{P}\mathbf{V}\mathbf{P}' = \sigma^2\mathbf{I}_n,\end{aligned}$$

  其中，最后一个等号成立是因为 $\mathbf{P}\mathbf{V}\mathbf{P}'$ 的逆矩阵是 $\mathbf{I}_n$。
- 因为 $\tilde{\boldsymbol{\varepsilon}}$ 是 $\boldsymbol{\varepsilon}$ 的线性变换，并且 $\mathbf{P}$ 是 $\mathbf{X}$ 的函数，所以，若 $\boldsymbol{\varepsilon} \mid \mathbf{X}$ 是正态分布，则 $\tilde{\boldsymbol{\varepsilon}} \mid \tilde{\mathbf{X}}$ 也是正态分布。

因此，转换后的变量满足古典线性回归模型的所有假设，对式 (2.31) 进行最小二乘估计便可以得到广义最小二乘估计量，记为 $\hat{\boldsymbol{\beta}}_{gls}$。关于广义最小二乘估计量，我们有如下定理：

定理 2.8. 广义最小二乘的小样本性质

(a) 广义最小二乘估计量 $\hat{\boldsymbol{\beta}}_{gls}$ 的解析表达式为：

$$\hat{\boldsymbol{\beta}}_{gls} = (\tilde{\mathbf{X}}'\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}'\tilde{\mathbf{y}} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}.$$

(b) (无偏性) 若假设 2.1-2.3 成立，$\mathbb{E}(\hat{\boldsymbol{\beta}}_{gls}|\mathbf{X}) = \boldsymbol{\beta}$；

(c) 若假设 2.1-2.3 和式 (2.30) 成立，$\mathrm{Var}(\hat{\boldsymbol{\beta}}_{gls}|\mathbf{X}) = \sigma^2 \cdot (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}$；

(d) (GLS 的有效性) 若假设 2.1-2.3 和式 (2.30) 成立，GLS 估计量是有效的，即，对任意的关于 $\mathbf{y}$ 线性并且无偏的估计量的条件方差都在矩阵意义上大于或等

于 $\mathrm{Var}(\hat{\boldsymbol{\beta}}_{gls}|\mathbf{X})$。 ♡

证明 (a) 因为 GLS 估计量是对式 (2.31) 做最小二乘估计，所以，由式 (2.8)，

$$\begin{aligned}\hat{\boldsymbol{\beta}}_{gls} &= (\tilde{\mathbf{X}}'\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}'\tilde{\mathbf{y}} \\ &= \left((\mathbf{PX})'(\mathbf{PX})\right)^{-1}(\mathbf{PX})'\mathbf{Py} \\ &= \left(\mathbf{X}'\mathbf{P}'\mathbf{PX}\right)^{-1}\left(\mathbf{X}'\mathbf{P}'\mathbf{Py}\right) \\ &= \left(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}.\end{aligned}$$

(b) 容易验证：$\hat{\boldsymbol{\beta}}_{gls}-\boldsymbol{\beta}=(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\boldsymbol{\varepsilon}$，因此，

$$\begin{aligned}\mathbb{E}(\hat{\boldsymbol{\beta}}_{gls}-\boldsymbol{\beta}|\mathbf{X}) &= \mathbb{E}\left[(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\boldsymbol{\varepsilon}|\mathbf{X}\right] \\ &= (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbb{E}\left[\boldsymbol{\varepsilon}|\mathbf{X}\right]=\mathbf{0}.\end{aligned}$$

(c) 利用 (b) 中 $\hat{\boldsymbol{\beta}}_{gls}-\boldsymbol{\beta}$ 的表达式，我们有：

$$\begin{aligned}\mathrm{Var}(\hat{\boldsymbol{\beta}}_{gls}|\mathbf{X}) &= \mathrm{Var}(\hat{\boldsymbol{\beta}}_{gls}-\boldsymbol{\beta}|\mathbf{X}) \\ &= (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathrm{Var}(\boldsymbol{\varepsilon}|\mathbf{X})\mathbf{V}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1} \\ &= \sigma^2\cdot(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1},\end{aligned}$$

其中，最后一个等号成立是因为式 (2.30) 成立。

最后，因为式 (2.31) 满足古典线性回归模型的所有假设，对其做最小二乘估计得到的估计量（即 GLS 估计量）是最有效的，(d) 得证。 □

GLS 小样本性质的成立需要严格外生条件，我们已经知道严格外生性不适用于时间序列数据，因此，我们通常不使用 GLS 修正误差项自相关的影响。

加权最小二乘

本小节讨论 GLS 的一个特例：加权最小二乘（Weighted Least Squares, WLS）。假设以 $\mathbf{X}$ 为条件误差项之间不存在相关性，则 $\mathbf{V}$ 是对角阵。令 $v_i(\mathbf{X})$ 是 $\mathbf{V}(\mathbf{X})$ 对角线上的第 i 个元素，我们有：

$$\mathbb{E}(\varepsilon_i^2\mid\mathbf{X})=\sigma^2\cdot v_i(\mathbf{X}).$$

此外，容易验证 $\mathbf{P}$ 此时也为对角阵，其对角线上第 i 个元素是 $1/\sqrt{v_i(\mathbf{X})}$。因此，经过 $\mathbf{P}$ 变换后的数据 $(\tilde{\mathbf{y}},\tilde{\mathbf{X}})$ 为：

$$\tilde{y}_i=\frac{y_i}{\sqrt{v_i(\mathbf{X})}},\quad \tilde{\mathbf{x}}_i=\frac{\mathbf{x}_i}{\sqrt{v_i(\mathbf{X})}}\quad (i=1,2,\ldots,n).$$

于是，若条件异方差的形式已知（即 $v_i(\mathbf{X})$ 已知），广义最小二乘估计量可以通过先将第 i 个观测的数据除以 $\sqrt{v_i(\mathbf{X})}$ 进行数据变换，再对变换后的数据进行 OLS 估计得到。此估计方法称为加权最小二乘。

若样本 $\{y_i,\mathbf{x}_i\}_{i=1}^n$ 满足独立同分布（independent and identically distributed，i.i.d.），则误差项满足无条件同方差（$\mathbb{E}(\varepsilon_i^2)$ 不依赖于 i），但是，误差项可能存在条件异方差。在 i.i.d. 的假设下，$\mathbb{E}(\varepsilon_i^2\mid\mathbf{X})$ 只与 $\mathbf{x}_i$ 有关，不同观测 i 的 $E(\varepsilon_i^2\mid\mathbf{x}_i)$ 具有相同的函数形式，

即，

$$v_i(\mathbf{X}) = v(\mathbf{x}_i), \quad (i = 1, 2, \ldots, n).$$

因此，对于 i.i.d. 的样本，做加权最小二乘估计仅需要知道函数 $v(\cdot)$。

可行广义最小二乘

使用 GLS 需要知道 $n \times n$ 维矩阵 $\mathbf{V}$ 的信息，虽然 $\mathbf{V}$ 是对称矩阵，仍然需要确定 $n(n+1)/2$ 个元素的取值。本节研究如何利用样本数据的信息对 $\mathbf{V}$ 进行估计，再根据 $\mathbf{V}$ 的估计结果做 GLS。值得一提的是，样本量为 n 的样本通常无法估计 $n(n+1)/2$ 个未知参数，我们需要对 $\mathbf{V}$ 的结构做一些假设。

为了讨论的方便，我们假设 $\{y_i, \mathbf{x}_i\}$ 是 i.i.d. 的。同时，假设：

$$\mathbb{E}(\varepsilon_i^2 \mid \mathbf{x}_i) = \mathbf{z}_i'\boldsymbol{\alpha}, \tag{2.32}$$

其中，$\mathbf{z}_i$ 是 $\mathbf{x}_i$ 的函数。通常，$\boldsymbol{\alpha}$ 是未知的，我们可以通过线性回归来估计 $\boldsymbol{\alpha}$。根据式 (2.32)，由 CEF 分解性质，我们有：

$$\varepsilon_i^2 = \mathbf{z}_i'\boldsymbol{\alpha} + \eta_i,$$

其中，$\eta_i \equiv \varepsilon_i^2 - \mathbb{E}(\varepsilon_i^2|\mathbf{x}_i)$。此外，$\eta_i$ 均值独立于 $\mathbf{z}_i$，事实上，

$$\mathbb{E}(\eta_i|\mathbf{z}_i) = \mathbb{E}\left\{\left[\varepsilon_i^2 - \mathbb{E}(\varepsilon_i^2|\mathbf{x}_i)\right]|\mathbf{z}_i\right\} = \mathbb{E}(\varepsilon_i^2|\mathbf{z}_i) - \mathbb{E}\left\{\mathbb{E}(\varepsilon_i^2|\mathbf{x}_i)|\mathbf{z}_i\right\} = 0,$$

其中，最后一个等号成立是利用了迭代期望律。因此，可以通过 ε_i^2 对 $\mathbf{z}_i$ 的回归估计 $\boldsymbol{\alpha}$。但是，ε_i^2 未知，可以考虑用残差的平方 e_i^2 代替 ε_i^2。于是，我们有如下的可行广义最小二乘（Feasible Generalized Least Squares, FGLS）：

步骤 1：利用 OLS 估计 $y_i = \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i$ 并计算残差 e_i；

步骤 2：将 e_i^2 对 $\mathbf{z}_i$ 进行回归得到最小二乘估计量 $\hat{\boldsymbol{\alpha}}$；

步骤 3：利用 WLS 重新估计 $y_i = \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i$，观测 i 对应的权重是 $1/\sqrt{\mathbf{z}_i'\hat{\boldsymbol{\alpha}}}$。

FGLS 的小样本性质不易推导，我们将在下一节讨论 FGLS 的大样本性质。此外，FGLS 的性质依赖于式 (2.32) 是否设置正确，如果存在条件方差形式的误设，FGLS 在大样本条件下的有效性可能不成立。式 (2.32) 虽然允许 $\mathbf{z}_i$ 是 $\mathbf{x}_i$ 的非线性函数，但是无法保证 $\mathbf{z}_i'\boldsymbol{\alpha}$ 始终大于 0。为了保证条件方差为正，一个常用的假设为：

$$\mathbb{E}(\varepsilon_i^2 \mid \mathbf{x}_i) = \sigma^2 \exp\left(\mathbf{z}_i'\boldsymbol{\alpha}\right). \tag{2.33}$$

上式也可以记为：

$$\varepsilon_i^2 = \sigma^2 \exp\left(\mathbf{z}_i'\boldsymbol{\alpha}\right) v_i,$$

其中，$\mathbb{E}(v_i|\mathbf{x}_i) = 1$。如果进一步假设 v_i 独立于 $\mathbf{x}_i$，我们有：

$$\begin{aligned}
\log(\varepsilon_i^2) &= \log(\sigma^2) + \mathbf{z}_i'\boldsymbol{\alpha} + \log(v_i) \\
&= \underbrace{\log(\sigma^2) + \alpha_1 + \mathbb{E}[\log(v_i)]}_{\alpha_0} + \mathbf{z}_{-1,i}'\boldsymbol{\alpha}_{-1} + \underbrace{\log(v_i) - \mathbb{E}[\log(v_i)]}_{e_i} \\
&\equiv \alpha_0 + \mathbf{z}_{-1,i}'\boldsymbol{\alpha}_{-1} + e_i,
\end{aligned}$$

其中，$\mathbf{z}_{-1,i}$ 是 $\mathbf{z}_i$ 中除常数项之外的变量；$\boldsymbol{\alpha}_{-1}$ 是 $\boldsymbol{\alpha}$ 中除常数项系数 α_1 之外的系数；e_i

的期望为 0 且独立于 $\mathbf{z}_i$。[9] 于是，针对形如式 (2.33) 的假设，我们可以通过如下 FGLS 过程估计回归系数 $\boldsymbol{\beta}$：

步骤 1*：利用 OLS 估计 $y_i = \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i$ 并计算残差 e_i；

步骤 2*：将 $\log(e_i^2)$ 对 $\mathbf{z}_i$ 进行回归得到拟合值 $\hat{g}_i$，并计算 $\hat{h}_i \equiv \exp(\hat{g}_i)$；

步骤 3*：利用 WLS 重新估计 $y_i = \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i$，观测 i 对应的权重是 $1/\sqrt{\hat{h}_i}$。

在步骤 2* 中，我们还可以将 $\log(e_i^2)$ 对步骤 1* 回归的拟合值 $\hat{y}_i$ 以及拟合值的平方 $\hat{y}_i^2$ 进行回归（回归中应包含常数项）得到 $\hat{g}_i$。关于 FGLS 更详细的讨论可以参考 Wooldridge (2019) 的第八章。

FGLS 的 Stata 实现

我们通过一个例子来介绍 FGLS 在 Stata 中如何实现。我们考虑互联网信息搜索与股票系统风险（systemic risk）之间的关系。被解释变量为股票系统风险（crash），是根据 Chabi-Yo et al. (2018) 中的混合 Copula 方法估计的股票随大盘暴跌的条件概率；核心解释变量是股票互联网信息搜索强度（IS），使用股票互联网信息搜索指数的对数一阶差分进行测度；回归中的控制变量有：股票周收益率的年度平均（wret）、超额换手率（本年度周均换手率与上年度周均换手率之差，wturn）、资产负债率（总负债/总资产，lev）、资产收益率（净利润/总资产，roa）、第一大股东持股比例（top）以及董事长和总经理两职合一的虚拟变量[10]（conp）。互联网信息搜索指数数据来自中国研究数据服务平台（CNRDS）中的网络搜索指数数据库（WSVI），上市公司信息、财务数据和股票数据等均来自 CSMAR 数据库。解释变量的样本来自 2016 年我国 A 股上市公司，经过剔除金融和保险行业公司以及剔除 ST、*ST 和 PT 公司等处理后，我们最终的样本量为 2397，被解释变量是 2017 年样本上市公司的股票系统风险。本例使用的 Stata 代码如下：

```
. use FGLS_Eg.dta, replace
. quietly reg crash IS wret wturn lev roa top conp
. estat hettest      //Breusch-Pagan异方差检验
H0: Constant variance
    chi2(1) = 19.18
Prob > chi2 = 0.0000
. predict e, res     //求残差，保存在e中
. gen logesq=ln(e^2)
. predict haty       //求拟合值，保存在haty中
. gen haty2 = haty^2 //计算拟合值的平方
. quietly reg logesq haty haty2
  //$\log(e_{i}^{2})$对拟合值和拟合值的平方进行回归
. predict esqhat     //求拟合值，保存在esqhat中
```

[9] 虽然 $\mathbb{E}(v_i) = 1$，但是 $\mathbb{E}[\log(v_i)]$ 可能不为 0，这是因为求期望和求对数不能换序。所以我们在构造误差项 e_i 时从 $\log(v_i)$ 中减去了它的期望，这样可以保证 e_i 的期望为 0，只要回归里有常数项，这样做就是合理的。

[10] 如果董事长和总经理两职合一，则此变量取值为 1；否则，此变量取值为 0。

```
14 . gen h=exp(esqhat)  //计算$\hat{h}$
15 . reg crash IS wret wturn lev roa top conp [aw= 1/h]
16   //重新回归，以$1/\hat{h}$作为权重
17 . est sto AW          //保存回归结果
18   //*****************************************************//
19   //自定义权重，对数据进行转换
20 . gen w = 1/(h)^0.5  //注意用于转换数据的权重是$1/\sqrt{\hat{h}_{i}}$
21 . foreach i in crash IS wret wturn lev roa top conp {
22     gen ‘i’w=‘i’*w    //数据转换
23     }
24 . reg crashw ISw wretw wturnw levw roaw topw conpw w, noc
25   //因为常数项也要做数据转换，回归中加入的w对应转换前的常数项，
26   // “, noc” 表示回归中没有常数项
27 . est sto NOC
28   //*****************************************************//
29   //也可以利用第三方命令wls0或regwls
30   //如果未安装wls0和regwls，使用findit wls0寻找下载地址
31 . wls0 crash IS wret wturn lev roa top conp, wvar(haty haty2)
     type(loge2)
32 . est sto WLS0
33 . regwls crash IS wret wturn lev roa top conp, wvar(haty haty2)
     type(loge2)
34 . est sto REGWLS
```

在程序的第 3 行，我们做了 Breusch-Pagan 异方差检验[11]，检验结果拒绝同方差的原假设。我们利用步骤 2* 的方式计算 $\hat{h}_i$，其中 $\mathbf{z}_i$ 取为拟合值 $\hat{y}_i$ 以及拟合值的平方 $\hat{y}_i^2$。第 15 行程序利用 reg 命令中的 aw 对回归进行加权，回归结果对应表 2.2 中“AW”一列；第 20 行程序生成了对数据进行转换的权重 w，注意 aw 中使用的权重是 w 的平方；第 24 行程序对转换后的数据进行没有常数项的回归，回归结果对应表 2.2 中“NOC”一列；第 31 行和第 33 行程序分别利用非官方命令 wls0 和 regwls 进行 FGLS 估计，回归结果对应表 2.2 中以命令名称命名的列中。“AW”和“NOC”对除常数项以外的解释变量系数的估计结果是相同的；“WLS0”和“REGWLS”对所有系数的估计结果都是相同的，但是与“AW”的回归结果不同。两者区别在于“WLS0”和“REGWLS”在计算 $\hat{h}_i$ 时是对程序第 13 行得到的拟合值 esqhat 取平方。从回归结果看，IS 系数是显著为负的，说明关于上市公司的互联网信息搜索强度越大，其面临的系统风险越低。

[11] 我们将在第3章详细介绍异方差检验。

表 2.2: 各种 FGLS 实现方法的结果比较

	(1) OLS	(2) AW	(3) NOC	(4) WLS0	(5) REGWLS
IS	-0.0121**	-0.0117**		-0.0122**	-0.0122**
	(0.00496)	(0.00485)		(0.00499)	(0.00499)
ISw			-0.0117**		
			(0.00485)		
_cons	0.0570***	0.0602***		0.0561***	0.0561***
	(0.00623)	(0.00587)		(0.00635)	(0.00635)
R^2	0.029	0.031	0.557	0.028	0.028
adj. R^2	0.026	0.028	0.556	0.025	0.025
F	10.075	10.934	375.791	9.909	9.909
N	2397	2397	2397	2397	2397

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

2.6 大样本性质及假设检验

前面我们讨论了最小二乘估计量的小样本性质以及基于正态假设的统计推断理论，然而，小样本性质的成立需要严格外生性和条件同方差等强假设条件，正态假设对多数金融数据也未必适用。因此，本节我们放松古典线性回归模型的部分假设，研究最小二乘估计量的大样本性质。所谓大样本性质是指样本量 n 趋于无穷时成立的性质。阅读本节内容需要读者具备一定的渐近理论基础，附录 B 中的 B.4 提供了相关知识的简要介绍。

大样本性质的推导基于以下关于数据生成过程（Data Generating Process, DGP）的假设：

假设 2.1*：（线性假设）

$$y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_K x_{iK} + \varepsilon_i = \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i, \quad (i = 1, 2, \ldots, n),$$

其中，$\mathbf{x}_i = (x_{i1}, \ldots, x_{iK})'$；$\boldsymbol{\beta} = (\beta_1, \ldots, \beta_K)'$ 是待估计的未知参数；ε_i 是不可观测的误差项。

假设 2.2*：（遍历平稳）

$$(K+1)\text{-维向量随机过程 } \{y_i, \mathbf{x}_i\} \text{ 满足联合遍历平稳。}$$

假设 2.3*：（前定（predetermined）假设）所有解释变量都是前定的，即，解释变量与同期的误差项满足正交性：

$$\text{对所有的 } i \in \{1, \ldots, n\} \text{ 和 } k \in \{1, 2, \ldots, K\}, \quad \mathbb{E}(x_{ik}\varepsilon_i) = 0,$$

表示为向量形式为：

$$\mathbb{E}[\mathbf{x}_i \cdot (y_i - \mathbf{x}_i'\boldsymbol{\beta})] = \mathbf{0},$$

或 $\mathbb{E}(\mathbf{g}_i) = \mathbf{0}$，其中，$\mathbf{g}_i \equiv \mathbf{x}_i \cdot \varepsilon_i$。

假设 2.4*：（秩条件）

$$K \times K \text{ 维矩阵 } \boldsymbol{\Sigma}_{\mathbf{xx}} = \mathbb{E}(\mathbf{x}_i\mathbf{x}_i') \text{ 是有限的非奇异阵。}$$

假设 2.5*：（鞅差假设）

$$\{\mathbf{g}_i\} \text{ 是鞅差序列，} K \times K \text{ 维矩阵 } \mathbf{S} = \mathbb{E}(\mathbf{g}_i\mathbf{g}_i') \text{ 非奇异。}$$

关于假设的说明

因为本节中对古典假设的放松主要是针对时间序列数据，所以，我们可以把变量的下标 i 理解成时间。例如，在假设 2.2* 中，我们称 $\{y_i, \mathbf{x}_i\}_{i=1}^n$ 为随机过程；在假设 2.5* 中，我们要求 $\{\mathbf{g}_i\}$ 是鞅差序列。如果将 i 换成 t，这些表述较容易理解，我们这里沿用 i 的原因是为了本章的符号统一。

假设 2.2* 要求 $\{y_i, \mathbf{x}_i\}$ 是遍历平稳的，因此，$\varepsilon_i = y_i - \mathbf{x}_i'\boldsymbol{\beta}$ 也是遍历平稳的。由平稳性的定义，若 ε_i 存在有限的二阶矩 $\mathbb{E}(\varepsilon_i^2)$，则 $\mathbb{E}(\varepsilon_i^2)$ 对所有的 i 都相同，即，ε_i 满足无条件的同方差。但是，假设 2.2* 允许 ε_i 存在条件异方差。例如，假设 $\varepsilon_i = \sigma(\mathbf{x}_i)\eta_i$，其中，$\{\eta_i\}$ 满足 i.i.d.、独立于 $\mathbf{x}_i$ 且期望为 0。于是，我们有：

$$\mathbb{E}(\varepsilon_i^2) = \mathbb{E}[\sigma^2(\mathbf{x}_i)\eta_i^2] = \mathbb{E}[\sigma^2(\mathbf{x}_i)]\mathbb{E}(\eta_i^2),$$

其中，第二个等号成立用到了 $\{\eta_i\}$ 与 $\mathbf{x}_i$ 相互独立。由 $\mathbf{x}_i$ 的平稳性和 η_i 满足 i.i.d.，我们有 $\mathbb{E}(\varepsilon_i^2)$ 与 i 无关。但是，

$$\mathbb{E}(\varepsilon_i^2 \mid \mathbf{x}_i) = \mathbb{E}(\sigma^2(\mathbf{x}_i)\eta_i^2 \mid \mathbf{x}_i) = \sigma^2(\mathbf{x}_i)\mathbb{E}(\eta_i^2).$$

因此，$\mathbb{E}(\varepsilon_i^2 \mid \mathbf{x}_i)$ 随着 $\mathbf{x}_i$ 的变化而变化。

在假设 2.3* 中，我们放松了严格外生假设，要求 $\mathbb{E}(x_{ik}\varepsilon_i) = 0$。严格外生性假设对时间序列数据通常是不成立的，但是，假设 2.3* 适用于时间序列数据。“前定”假设指解释变量与同期的误差项正交，但是时间序列模型中的解释变量一般是被解释变量的滞后期，因此，“前定”假设要求 i 期误差项与滞后期的被解释变量正交。此假设的合理性就在于滞后期的被解释变量是“事先”确定的，与未来期的误差项不相关。例如，考虑如下一阶自回归模型：

$$y_i = \beta_0 + \beta_1 y_{i-1} + \varepsilon_i.$$

在此模型中 $x_{i2} = y_{i-1}$，$\mathbb{E}(x_{i2}\varepsilon_i) = 0$ 等价于 $\mathbb{E}(y_{i-1}\varepsilon_i) = 0$；同时，若回归有常数项，假设 2.3* 保证了 $\mathbb{E}(\varepsilon_i) = 0$。因此，假设 2.3* 实际上要求 y_{i-1} 和 ε_i 不相关。直观上，y_{i-1} 是 $i-1$ 期的观测数据，而 ε_i 是 i 期数据的随机扰动项，两者之间不存在相关性是符合常理的。

秩条件假设 2.4* 要求 $\mathbb{E}(\mathbf{x}_i\mathbf{x}_i')$ 是有限的非奇异矩阵，同时，由遍历定理 B.7，我们有：

$$\mathbf{S}_{\mathbf{xx}} \equiv \frac{1}{n}\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i' \underset{a.s.}{\to} \mathbb{E}(\mathbf{x}_i\mathbf{x}_i') = \boldsymbol{\Sigma}_{\mathbf{xx}}.$$

注意到 $\mathbf{S}_{\mathbf{xx}}$ 还可以表示为 $\mathbf{X}'\mathbf{X}/n$，于是，$\mathbf{X}'\mathbf{X}/n$ 依概率 1 收敛到 $\boldsymbol{\Sigma}_{\mathbf{xx}}$。因为 $\mathbf{X}'\mathbf{X}/n$ 非奇异当且仅当 $\operatorname{rank}(\mathbf{X}) = K$，所以，无完全共线性假设 2.3 渐近成立。

假设 2.5* 要求 $\{\mathbf{g}_i\}$ 是鞅差序列，由附录 B.4 中关于鞅差序列的定义和简单迭代期望律，我们有 $\mathbb{E}(\mathbf{g}_i) = \mathbf{0}$，因此，假设 2.5* 强于假设 2.3*。在下一小节我们将要看到，保证

最小二乘估计量一致性的核心假设是假设 2.3*，但是为了建立大样本下的统计推断理论，我们需要利用假设 2.5* 证明最小二乘估计量的渐近正态性。在小样本下，我们通过假设误差项服从正态分布推导出 t 统计量和 F 统计量的精确分布；在大样本下，我们放松了正态假设，通过最小二乘估计量的渐近正态性推出各个假设检验统计量的渐近分布。假设 2.5* 的一个充分条件是：

$$\mathbb{E}(\varepsilon_i \mid \varepsilon_{i-1}, \varepsilon_{i-2}, \ldots, \varepsilon_1, \mathbf{x}_i, \mathbf{x}_{i-1}, \ldots, \mathbf{x}_1) = 0. \tag{2.34}$$

在条件信息集中我们加入了滞后期的误差项以及同期和滞后期的解释变量。从此条件可以推出误差项无自相关，而且误差项与同期和滞后期的解释变量也不相关。

如果回归中有常数项，由假设 2.5* 可以得到：

$$\mathbb{E}(\varepsilon_i \mid \mathbf{g}_{i-1}, \mathbf{g}_{i-2}, \ldots, \mathbf{g}_1) = 0.$$

再由迭代期望律，我们有：

$$\mathbb{E}(\varepsilon_i \mid \varepsilon_{i-1}, \varepsilon_{i-2}, \ldots, \varepsilon_1) = 0.$$

因此，在有常数项的回归中，假设 2.5* 成立的必要条件是 $\{\varepsilon_i\}$ 是鞅差序列。鞅差序列的一个重要性质是它不存在自相关。事实上，对任意的整数 $j > 0$，我们有：

$$\begin{aligned}
\mathrm{Cov}(\mathbf{g}_i, \mathbf{g}_{i-j}) &= \mathbb{E}(\mathbf{g}_i \mathbf{g}_{i-j}') && (\text{因为 } \mathbb{E}(\mathbf{g}_i) = \mathbf{0}) \\
&= \mathbb{E}\left[\mathbb{E}\left(\mathbf{g}_i \mathbf{g}_{i-j}' \mid \mathbf{g}_{i-j}\right)\right] && (\text{由迭代期望律}) \\
&= \mathbb{E}\left[\mathbb{E}\left(\mathbf{g}_i \mid \mathbf{g}_{i-j}\right) \mathbf{g}_{i-j}'\right] && (\text{由定理 1.3}) \\
&= \mathbb{E}\left\{\mathbb{E}\left[\mathbb{E}\left(\mathbf{g}_i \mid \mathscr{F}_{i-1}\right) \mid \mathbf{g}_{i-j}\right] \mathbf{g}_{i-j}'\right\} && (\mathscr{F}_{i-1} \text{ 包含 } \mathbf{g}_{i-j}) \\
&= \mathbf{0}, && (\text{根据鞅差的定义，} \mathbb{E}\left(\mathbf{g}_i \mid \mathscr{F}_{i-1}\right) = \mathbf{0})
\end{aligned}$$

其中，$\mathscr{F}_{i-1} = (\mathbf{g}_{i-1}, \mathbf{g}_{i-2}, \ldots, \mathbf{g}_1)$。所以，若 $\{\varepsilon_i\}$ 存在自相关，则假设 2.5* 不成立。我们将在第 3 章讨论如何检验 $\{\varepsilon_i\}$ 是否存在自相关。

最小二乘估计的大样本性质

最小二乘的大样本性质是样本量 n 趋于无穷才成立的性质，为了理解大样本性质的含义，我们应该把最小二乘估计量 $\hat{\boldsymbol{\beta}}_{ols}$ 看成是与样本量 n 有关的随机变量。在大样本下，我们有如下结论：

定理 2.9. 最小二乘估计量的大样本性质

(a) （一致性）若假设 2.1*-2.4* 成立，$\operatorname*{plim}_{n\to\infty} \hat{\boldsymbol{\beta}}_{ols} = \boldsymbol{\beta}$；

(b) （渐近正态性）若将 (a) 中假设 2.3* 加强为假设 2.5*，则有：当 $n \to \infty$ 时，

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{ols} - \boldsymbol{\beta}) \underset{d}{\to} \mathcal{N}(\mathbf{0}, \mathrm{Avar}(\hat{\boldsymbol{\beta}}_{ols})), \tag{2.35}$$

其中，

$$\mathrm{Avar}(\hat{\boldsymbol{\beta}}_{ols}) = \boldsymbol{\Sigma}_{\mathbf{xx}}^{-1} \mathbf{S} \boldsymbol{\Sigma}_{\mathbf{xx}}^{-1}.$$

(c) 假设存在 $\mathbf{S}$ 的一致估计量 $\hat{\mathbf{S}}$，则，在假设 2.2* 成立的情况下，$\mathrm{Avar}(\hat{\boldsymbol{\beta}}_{ols})$ 的一

致估计量为：

$$\widehat{\mathrm{Avar}(\hat{\boldsymbol{\beta}}_{ols})} = \mathbf{S}_{\mathbf{xx}}^{-1}\hat{\mathbf{S}}\mathbf{S}_{\mathbf{xx}}^{-1},$$

其中，

$$\mathbf{S}_{\mathbf{xx}} \equiv \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i' = \frac{1}{n}\mathbf{X}'\mathbf{X}.$$

♡

证明 (a) 由式 (2.17)，我们有：

$$\begin{aligned}\hat{\boldsymbol{\beta}}_{ols} - \boldsymbol{\beta} &= \left(\frac{1}{n}\mathbf{X}'\mathbf{X}\right)^{-1}\left(\frac{1}{n}\mathbf{X}'\boldsymbol{\varepsilon}\right) = \left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i'\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\cdot\varepsilon_i\right) \\ &\equiv \mathbf{S}_{\mathbf{xx}}^{-1}\bar{\mathbf{g}}, \end{aligned} \tag{2.36}$$

其中，$\bar{\mathbf{g}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\cdot\varepsilon_i$。由假设 2.2*，$\{\mathbf{x}_i\mathbf{x}_i'\}$ 是遍历平稳的，所以，由遍历定理 B.7，$\mathbf{S}_{\mathbf{xx}} \underset{p}{\to} \boldsymbol{\Sigma}_{\mathbf{xx}}$。再由假设 2.4*，$\boldsymbol{\Sigma}_{\mathbf{xx}}$ 是可逆的，根据连续映射定理 B.3，[12] 我们有：$\mathbf{S}_{\mathbf{xx}}^{-1} \underset{p}{\to} \boldsymbol{\Sigma}_{\mathbf{xx}}^{-1}$。

类似地，$\bar{\mathbf{g}} \underset{p}{\to} \mathbb{E}(\mathbf{g}_i) = \mathbf{0}$，于是，再由连续映射定理 B.3，$\mathbf{S}_{\mathbf{xx}}^{-1}\bar{\mathbf{g}} \underset{p}{\to} \boldsymbol{\Sigma}_{\mathbf{xx}}^{-1}\mathbf{0} = \mathbf{0}$。所以，$\operatorname*{plim}_{n\to\infty}\hat{\boldsymbol{\beta}}_{ols} = \boldsymbol{\beta}$。

(b) 由式 (2.36)，

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{ols} - \boldsymbol{\beta}) = \mathbf{S}_{\mathbf{xx}}^{-1}(\sqrt{n}\bar{\mathbf{g}}).$$

由假设 2.5* 和遍历平稳鞅差序列的中心极限定理 B.8，

$$\sqrt{n}\bar{\mathbf{g}} \underset{d}{\to} \mathcal{N}(\mathbf{0}, \mathbf{S}).$$

于是，由 Slutsky 定理 B.4，$\sqrt{n}(\hat{\boldsymbol{\beta}}_{ols} - \boldsymbol{\beta}) \underset{d}{\to} \mathcal{N}(\mathbf{0}, \mathrm{Avar}(\hat{\boldsymbol{\beta}}_{ols}))$。

(c) 因为 $\mathbf{S}_{\mathbf{xx}} \underset{p}{\to} \boldsymbol{\Sigma}_{\mathbf{xx}}$ 和 $\hat{\mathbf{S}} \underset{p}{\to} \mathbf{S}$，由连续映射定理 B.3，我们有：$\widehat{\mathrm{Avar}(\hat{\boldsymbol{\beta}}_{ols})} \underset{p}{\to} \mathrm{Avar}(\hat{\boldsymbol{\beta}}_{ols})$。 □

定理 2.9 的结论 (a) 告诉我们最小二乘估计量是一致的，假设 2.3* 在一致性的证明中起到关键作用。值得说明的是，虽然严格外生性假设 2.2 强于假设 2.3*，但是由假设 2.2 保证的最小二乘估计量的无偏性和一致性没有强弱之分，也就是说，无偏性推不出一致性，当然一致性也推不出无偏性。我们通常说无偏性是一个很强的结论是指它对任意的 n ($\geqslant K$) 都成立，并非指它是比一致性更强的结论。例如，假设 $\mathrm{x}_i \overset{i.i.d.}{\sim} \mathcal{N}(\mu, \sigma^2)$，$i = 1, \ldots, n$。我们首先考虑期望 μ 的估计量 $\hat{\mu}_n = \mathrm{x}_1$，因为 $\mathbb{E}(\hat{\mu}_n) = \mathbb{E}(\mathrm{x}_1) = \mu$，所以 $\hat{\mu}_n$ 是无偏估计量；但是，$\mathrm{x}_1 \sim \mathcal{N}(\mu, \sigma^2)$，因此，$\hat{\mu}_n$ 不是一致估计量。我们再考虑方差 σ^2 的估计量，我们知道 $\frac{1}{n}\sum_{i=1}^{n}(\mathrm{x}_i - \bar{\mathrm{x}})^2$ 是 σ^2 的一致估计量，但是它不是 σ^2 的无偏估计量。

在定理 2.9 的 (c) 中，我们假设存在 $\mathbf{S}$ 的一致估计量 $\hat{\mathbf{S}}$，因为，

$$\mathbf{S} = \mathbb{E}(\mathbf{g}_i\mathbf{g}_i') = \mathbb{E}(\varepsilon_i^2\mathbf{x}_i\mathbf{x}_i'),$$

[12] 矩阵求逆是连续函数，参见 Bredon (2013) 第 53 页。

一个自然的想法是令

$$\hat{\mathbf{S}} \equiv \frac{1}{n}\sum_{i=1}^{n} e_i^2 \mathbf{x}_i \mathbf{x}_i',$$

即用残差 e_i 代替不可观测的 ε_i。为了证明 $\hat{\mathbf{S}}$ 是 $\mathbf{S}$ 的一致估计量，我们需要如下关于解释变量四阶矩的假设：

假设 2.6*：（解释变量四阶矩有限）对所有的 $k, j = 1, 2, \ldots, K$，$\mathbb{E}[(x_{ik}x_{ij})^2]$ 存在且有限。

定理 2.10. S 的一致估计

假设 $\mathbf{S} = \mathbb{E}(\mathbf{g}_i\mathbf{g}_i')$ 存在且有限。若假设 2.1*，2.2* 和 2.6* 成立，则 $\hat{\mathbf{S}}$ 是 $\mathbf{S}$ 的一致估计量。 ♡

证明 注意到

$$e_i = y_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}}_{ols} = y_i - \mathbf{x}_i'\boldsymbol{\beta} - \mathbf{x}_i'(\hat{\boldsymbol{\beta}}_{ols} - \boldsymbol{\beta}) = \varepsilon_i - \mathbf{x}_i'(\hat{\boldsymbol{\beta}}_{ols} - \boldsymbol{\beta}),$$

于是，

$$e_i^2 = \varepsilon_i^2 - 2(\hat{\boldsymbol{\beta}}_{ols} - \boldsymbol{\beta})'\mathbf{x}_i \cdot \varepsilon_i + (\hat{\boldsymbol{\beta}}_{ols} - \boldsymbol{\beta})'\mathbf{x}_i\mathbf{x}_i'(\hat{\boldsymbol{\beta}}_{ols} - \boldsymbol{\beta}). \tag{2.37}$$

我们考虑矩阵 $\hat{\mathbf{S}}$ 在 (l, m) 位置上的元素，$(l, m = 1, 2, \ldots, K)$，我们有：

$$\begin{aligned}
\hat{\mathbf{S}}_{lm} &\equiv \frac{1}{n}\sum_{i=1}^{n} e_i^2 x_{il}x_{im} \\
&= \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i^2 x_{il}x_{im} - 2(\hat{\boldsymbol{\beta}}_{ols} - \boldsymbol{\beta})'\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i \cdot x_{il}x_{im}\varepsilon_i\right) \\
&\quad + (\hat{\boldsymbol{\beta}}_{ols} - \boldsymbol{\beta})'\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i' x_{il}x_{im}\right)(\hat{\boldsymbol{\beta}}_{ols} - \boldsymbol{\beta}).
\end{aligned}$$

由柯西-施瓦茨不等式 (A.1)、假设 2.6* 和假设 2.2*，$\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i \cdot x_{il}x_{im}\varepsilon_i$ 依概率收敛到有限值，例如，考虑此向量的第 k 个元素 $\frac{1}{n}\sum_{i=1}^{n} x_{ik} \cdot x_{il}x_{im}\varepsilon_i$，我们有：

$$\mathbb{E}(x_{ik} \cdot x_{il}x_{im}\varepsilon_i) \leqslant \mathbb{E}(|x_{ik} \cdot x_{il}x_{im}\varepsilon_i|) \leqslant \sqrt{\mathbb{E}(x_{ik}^2 x_{il}^2)\mathbb{E}(x_{im}^2\varepsilon_i^2)} < \infty,$$

其中，$\mathbb{E}(x_{ik}^2 x_{il}^2) < \infty$ 我们用到了假设 2.6*，而 $\mathbb{E}(x_{im}^2\varepsilon_i^2) < \infty$ 我们用到了 $\mathbf{S}$ 存在且有限。所以，由遍历定理 B.7 可知 $\frac{1}{n}\sum_{i=1}^{n} x_{ik} \cdot x_{il}x_{im}\varepsilon_i$ 收敛到 $\mathbb{E}(x_{ik} \cdot x_{il}x_{im}\varepsilon_i)$。同理，$\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i' x_{il}x_{im}$ 也依概率收敛到有限值。又因为 $\hat{\boldsymbol{\beta}}_{ols}$ 是一致估计量，$\hat{\boldsymbol{\beta}}_{ols} - \boldsymbol{\beta}$ 依概率收敛到 $\mathbf{0}$，于是，由连续映射定理 B.3 和遍历定理 B.7，

$$\hat{\mathbf{S}}_{lm} \underset{p}{\to} \mathbb{E}(\varepsilon_i^2 x_{il}x_{im}) = \mathbf{S}_{lm},$$

即，$\hat{\mathbf{S}}$ 的 (l, m) 位置上的元素依概率收敛到 $\mathbf{S}$ 的 (l, m) 位置上的元素，则有 $\hat{\mathbf{S}} \underset{p}{\to} \mathbf{S}$。□

大样本统计推断

多元线性回归的大样本统计推断基于在原假设成立的情况下检验统计量的渐近分布。我们首先讨论线性假设检验，有如下结论：

定理 2.11. 稳健的 t 统计量和 Wald 统计量

若假设 2.1*-2.6* 成立，则有，

(a) 在原假设 $\mathbb{H}_0: \beta_k = \beta_k^0$ 成立的情况下，

$$t_k \equiv \frac{\sqrt{n}(\hat{\beta}_{k,ols} - \beta_k^0)}{\sqrt{\widehat{\mathrm{Avar}}(\hat{\beta}_{k,ols})}} = \frac{\hat{\beta}_{k,ols} - \beta_k^0}{SE^*(\hat{\beta}_{k,ols})} \underset{d}{\to} \mathcal{N}(0,1), \tag{2.38}$$

其中，$\widehat{\mathrm{Avar}}(\hat{\beta}_{k,ols})$ 是 $\widehat{\mathrm{Avar}}(\hat{\boldsymbol{\beta}}_{ols})$ 第 (k,k) 位置上的元素，并且

$$SE^*(\hat{\beta}_{k,ols}) \equiv \sqrt{\frac{1}{n} \cdot \widehat{\mathrm{Avar}}(\hat{\beta}_{k,ols})} = \sqrt{\frac{1}{n} \cdot \left(\mathbf{S}_{\mathbf{xx}}^{-1}\hat{\mathbf{S}}\mathbf{S}_{\mathbf{xx}}^{-1}\right)_{kk}}. \tag{2.39}$$

$SE^*(\hat{\beta}_{k,ols})$ 称为异方差一致标准误（heteroskedasticity-consistent standard error），或异方差稳健标准误（heteroskedasticity-robust standard error），或 White 标准误；

(b) 令 $\mathbf{R}$ 是 $J \times K$ 维的行满秩矩阵，在原假设 $\mathbb{H}_0: \mathbf{R}\boldsymbol{\beta} = \mathbf{r}$ 成立的情况下，

$$W \equiv n \cdot (\mathbf{R}\hat{\boldsymbol{\beta}}_{ols} - \mathbf{r})' \left\{\mathbf{R}[\widehat{\mathrm{Avar}}(\hat{\boldsymbol{\beta}}_{ols})]\mathbf{R}'\right\}^{-1} (\mathbf{R}\hat{\boldsymbol{\beta}}_{ols} - \mathbf{r}) \underset{d}{\to} \chi^2(J). \tag{2.40}$$

♡

证明 (a) 由式 (2.35)，

$$\sqrt{n}(\hat{\beta}_{k,ols} - \beta_k^0) \underset{d}{\to} \mathcal{N}(0, \mathrm{Avar}(\hat{\beta}_{k,ols})).$$

同时，由定理 2.9 的 (c)，我们有：

$$\widehat{\mathrm{Avar}}(\hat{\beta}_{k,ols}) \underset{p}{\to} \mathrm{Avar}(\hat{\beta}_{k,ols}).$$

于是，由 Slutsky 定理 B.4，

$$t_k \equiv \frac{\sqrt{n}(\hat{\beta}_{k,ols} - \beta_k^0)}{\sqrt{\widehat{\mathrm{Avar}}(\hat{\beta}_{k,ols})}} \underset{d}{\to} \mathcal{N}(0,1).$$

(b) 注意到，W 可以表示为：

$$W = \mathbf{c}_n'\mathbf{Q}_n^{-1}\mathbf{c}_n, \quad \text{其中} \quad \mathbf{c}_n \equiv \sqrt{n}(\mathbf{R}\hat{\boldsymbol{\beta}}_{ols} - \mathbf{r}), \mathbf{Q}_n \equiv \mathbf{R}\widehat{\mathrm{Avar}}(\hat{\boldsymbol{\beta}}_{ols})\mathbf{R}'.$$

在 $\mathbb{H}_0$ 成立的情况下，$\mathbf{c}_n = \mathbf{R}\sqrt{n}(\hat{\boldsymbol{\beta}}_{ols} - \boldsymbol{\beta})$。因此，由式 (2.35)，

$$\mathbf{c}_n \underset{d}{\to} \mathbf{c} \quad \text{其中} \quad \mathbf{c} \sim \mathcal{N}(\mathbf{0}, \mathbf{R}\mathrm{Avar}(\hat{\boldsymbol{\beta}}_{ols})\mathbf{R}').$$

此外，

$$\mathbf{Q}_n \underset{p}{\to} \mathbf{Q} \quad \text{其中} \quad \mathbf{Q} \equiv \mathbf{R}\mathrm{Avar}(\hat{\boldsymbol{\beta}}_{ols})\mathbf{R}'.$$

因为 $\mathbf{R}$ 行满秩并且 $\mathrm{Avar}(\hat{\boldsymbol{\beta}}_{ols})$ 正定，所以 $\mathbf{Q}$ 可逆。于是，由 (B.8)，我们有：

$$W \underset{d}{\to} \mathbf{c}'\mathbf{Q}^{-1}\mathbf{c}.$$

因为 $J \times 1$ 维随机向量 $\mathbf{c}$ 服从正态分布，期望是 $\mathbf{0}$，协方差矩阵是 $\mathbf{Q}$，所以，

$$\mathbf{c}'\mathbf{Q}^{-1}\mathbf{c} \sim \chi^2(J).$$

□

在前文我们提到，满足假设 2.1*-2.6* 的 DGP 的误差项是可以存在条件异方差的，但是在假设 2.5* 下，误差项不存在自相关。因此，定理 2.11 中给出的标准误 (2.39) 适用于

误差项存在条件异方差的情况，但是不能用于误差项存在自相关的情况。

基于异方差稳健 t 统计量的假设检验的具体步骤如下：

1. 根据原假设 $\mathbb{H}_0$ 中参数的具体取值 β_k^0 计算由式 (2.38) 定义的 t 统计量；
2. 通过统计软件计算标准正态分布的临界值 $t_{\alpha/2}$（也可以通过查表的方式找到临界值），$t_{\alpha/2}$ 满足标准正态分布在此临界值右侧的阴影面积为 $\alpha/2$；
3. 如果第一步计算的 t_k 满足 $|t_k| < t_{\alpha/2}$，则接受 $\mathbb{H}_0$；否则，拒绝 $\mathbb{H}_0$。

基于 Wald 统计量 W 的假设检验的具体步骤如下：

1. 根据式 (2.40) 计算 Wald 统计量；
2. 通过统计软件计算 $\chi^2(J)$ 的临界值 $\chi^2_\alpha(J)$（也可以通过查表的方式找到临界值），$\chi^2_\alpha(J)$ 满足自由度为 J 的卡方分布在此临界值右侧的阴影面积为 α；
3. 如果第一步计算的 W 统计量满足 $W < \chi^2_\alpha(J)$，则接受 $\mathbb{H}_0$；否则，拒绝 $\mathbb{H}_0$。

与小样本统计推断不同，基于渐近分布的假设检验的真实显著性水平随着 n 的增大会越来越接近名义的显著性水平 α，而对固定的 n，真实的显著性水平一般不等于 α。

在大样本下，我们还可以检验关于 $\boldsymbol{\beta}$ 的非线性原假设 $\mathbb{H}_0 : \mathbf{a}(\boldsymbol{\beta}) = \mathbf{0}$，其中，$\mathbf{a}(\cdot)$ 是 $J \times 1$ 维的向量值函数且具有连续的一阶导。我们有如下结论：

定理 2.12. β 的非线性假设检验

令 $\mathbf{a}(\boldsymbol{\beta})$ 是 $J \times 1$ 维的向量值函数，具有连续的一阶导，记 $\mathbf{a}(\boldsymbol{\beta})$ 的一阶导在 $\boldsymbol{\beta}$ 处的取值为 $\mathbf{A}(\boldsymbol{\beta}) = \partial \mathbf{a}(\boldsymbol{\beta})/\partial \boldsymbol{\beta}'$。若假设 2.1*-2.6* 成立，且 $\mathbf{A}(\boldsymbol{\beta})$ 行满秩，在原假设 $\mathbb{H}_0 : \mathbf{a}(\boldsymbol{\beta}) = \mathbf{0}$ 成立的情况下，我们有：

$$W \equiv n \cdot \mathbf{a}(\hat{\boldsymbol{\beta}}_{ols})' \left\{ \mathbf{A}(\hat{\boldsymbol{\beta}}_{ols}) \widehat{\mathrm{Avar}(\hat{\boldsymbol{\beta}}_{ols})} \mathbf{A}(\hat{\boldsymbol{\beta}}_{ols})' \right\}^{-1} \mathbf{a}(\hat{\boldsymbol{\beta}}_{ols}) \underset{d}{\to} \chi^2(J).$$

♡

证明 由式 (2.35) 和 Delta 方法 B.5，我们有：

$$\sqrt{n}\left[\mathbf{a}(\hat{\boldsymbol{\beta}}_{ols}) - \mathbf{a}(\boldsymbol{\beta})\right] \underset{d}{\to} \mathbf{c}, \quad \mathbf{c} \sim \mathcal{N}(\mathbf{0}, \mathbf{A}(\boldsymbol{\beta})\mathrm{Avar}(\hat{\boldsymbol{\beta}}_{ols})\mathbf{A}(\boldsymbol{\beta})').$$

若 $\mathbb{H}_0$ 成立，则有：

$$\sqrt{n}\mathbf{a}(\hat{\boldsymbol{\beta}}_{ols}) \underset{d}{\to} \mathbf{c}, \quad \mathbf{c} \sim \mathcal{N}(\mathbf{0}, \mathbf{A}(\boldsymbol{\beta})\mathrm{Avar}(\hat{\boldsymbol{\beta}}_{ols})\mathbf{A}(\boldsymbol{\beta})').$$

因为 $\hat{\boldsymbol{\beta}}_{ols} \underset{p}{\to} \boldsymbol{\beta}$，由连续映射定理 B.3，$\mathbf{A}(\hat{\boldsymbol{\beta}}_{ols}) \underset{p}{\to} \mathbf{A}(\boldsymbol{\beta})$。因此，

$$\mathbf{A}(\hat{\boldsymbol{\beta}}_{ols})\widehat{\mathrm{Avar}(\hat{\boldsymbol{\beta}}_{ols})}\mathbf{A}(\hat{\boldsymbol{\beta}}_{ols})' \underset{p}{\to} \mathbf{A}(\boldsymbol{\beta})\mathrm{Avar}(\hat{\boldsymbol{\beta}}_{ols})\mathbf{A}(\boldsymbol{\beta})' = \mathrm{Var}(\mathbf{c}).$$

又因此 $\mathbf{A}(\boldsymbol{\beta})$ 行满秩且 $\mathrm{Avar}(\hat{\boldsymbol{\beta}}_{ols})$ 正定，所以，$\mathrm{Var}(\mathbf{c})$ 可逆。于是，

$$\sqrt{n}\mathbf{a}(\hat{\boldsymbol{\beta}}_{ols})' \left\{ \mathbf{A}(\hat{\boldsymbol{\beta}}_{ols})\widehat{\mathrm{Avar}(\hat{\boldsymbol{\beta}}_{ols})}\mathbf{A}(\hat{\boldsymbol{\beta}}_{ols})' \right\}^{-1} \sqrt{n}\mathbf{a}(\hat{\boldsymbol{\beta}}_{ols}) \underset{d}{\to} \mathbf{c}'\mathrm{Var}(\mathbf{c})^{-1}\mathbf{c} \sim \chi^2(J).$$

□

需要指出的是 Wald 统计量在固定样本量 n 时的真实显著性水平受原假设的函数形式影响。若原假设既可以表示成线性形式又可以表示成非线性形式，我们要尽可能的使用线性形式进行假设检验。例如：若想要检验的原假设为 $\mathbb{H}_0 : \frac{\beta_1}{\beta_2} = \theta_0$，我们应该将其转换为 $\mathbb{H}_0 : \beta_1 - \theta_0\beta_2 = 0$。

同方差假设下的大样本统计推断

为了与小样本的统计推断理论进行比较，本节讨论同方差假设下的大样本统计推断。为此，我们施加如下假设：

假设 2.7*：（条件同方差） $\mathbb{E}(\varepsilon_i^2 \mid \mathbf{x}_i) = \sigma^2 > 0$。

我们首先给出 σ^2 的一个一致估计量：

定理 2.13. σ^2 的一致估计

令 $e_i \equiv y_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}}_{ols}$ 是观测 i 的 OLS 残差。若假设 2.1*-2.4* 成立，

$$s^2 \equiv \frac{1}{n-K}\sum_{i=1}^{n} e_i^2 \underset{p}{\to} \mathbb{E}(\varepsilon_i^2) = \sigma^2.$$

♡

证明 因为

$$s^2 = \frac{n}{n-K}\left(\frac{1}{n}\sum_{i=1}^{n} e_i^2\right),$$

所以我们只需证明 e_i^2 的样本均值依概率收敛到 $\mathbb{E}(\varepsilon_i^2)$。对式 (2.37) 等号两端取样本均值得到：

$$\frac{1}{n}\sum_{i=1}^{n} e_i^2 = \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i^2 - 2(\hat{\boldsymbol{\beta}}_{ols} - \boldsymbol{\beta})'\bar{\mathbf{g}} + (\hat{\boldsymbol{\beta}}_{ols} - \boldsymbol{\beta})'\mathbf{S}_{\mathbf{xx}}(\hat{\boldsymbol{\beta}}_{ols} - \boldsymbol{\beta}).$$

由最小二乘估计量的一致性、假设 2.2*-2.3* 和遍历定理 B.7 易知：

$$\operatorname{plim}\left(\frac{1}{n}\sum_{i=1}^{n} e_i^2\right) = \operatorname{plim}\left(\frac{1}{n}\sum_{i=1}^{n} \varepsilon_i^2\right) = \mathbb{E}(\varepsilon_i^2).$$

□

有了 σ^2 的一致估计，我们可以证明如下结论：

定理 2.14. 同方差假设下的大样本假设检验

若假设 2.1*-2.5* 和假设 2.7* 成立，则有：

(a) 最小二乘估计量 $\hat{\boldsymbol{\beta}}_{ols}$ 一致且渐近服从正态分布，其渐近方差为 $\operatorname{Avar}(\hat{\boldsymbol{\beta}}_{ols}) = \sigma^2\boldsymbol{\Sigma}_{\mathbf{xx}}^{-1}$；

(b) $\operatorname{Avar}(\hat{\boldsymbol{\beta}}_{ols})$ 的一致估计量为 $\widehat{\operatorname{Avar}(\hat{\boldsymbol{\beta}}_{ols})} = s^2\mathbf{S}_{\mathbf{xx}}^{-1} = n \cdot s^2 \cdot (\mathbf{X}'\mathbf{X})^{-1}$；

(c) 在原假设 $\mathbb{H}_0 : \beta_k = \beta_k^0$ 下，小样本 t 统计量渐近服从标准正态分布；在原假设 $\mathbb{H}_0 : \mathbf{R}\boldsymbol{\beta} = \mathbf{r}$ 下，$J \cdot F$ 渐近服从 $\chi^2(J)$，其中，F 是小样本 F 统计量，J 是 $\mathbb{H}_0$ 中的约束个数。

♡

证明 (a) 若假设 2.7* 成立，我们有：

$$\mathbf{S} = \mathbb{E}(\mathbf{x}_i\mathbf{x}_i'\varepsilon_i^2) = \mathbb{E}[\mathbf{x}_i\mathbf{x}_i'\mathbb{E}(\varepsilon_i^2 \mid \mathbf{x}_i)] = \sigma^2\boldsymbol{\Sigma}_{\mathbf{xx}}.$$

所以，$\operatorname{Avar}(\hat{\boldsymbol{\beta}}_{ols}) = \boldsymbol{\Sigma}_{\mathbf{xx}}^{-1}\mathbf{S}\boldsymbol{\Sigma}_{\mathbf{xx}}^{-1} = \sigma^2\boldsymbol{\Sigma}_{\mathbf{xx}}^{-1}$。

(b) 由遍历定理 B.7，$\mathbf{S}_{\mathbf{xx}} \underset{p}{\to} \boldsymbol{\Sigma}_{\mathbf{xx}}$，所以，$\mathbf{S}_{\mathbf{xx}}^{-1} \underset{p}{\to} \boldsymbol{\Sigma}_{\mathbf{xx}}^{-1}$。再由定理 2.13，$s^2$ 是 σ^2 的一致估计量，于是，我们有：$s^2\mathbf{S}_{\mathbf{xx}}^{-1} \underset{p}{\to} \sigma^2\boldsymbol{\Sigma}_{\mathbf{xx}}^{-1}$。

(c) 小样本 t 统计量为

$$t_k \equiv \frac{\hat{\beta}_{k,ols} - \beta_k^0}{\sqrt{s^2 \cdot \left((\mathbf{X}'\mathbf{X})^{-1}\right)_{kk}}}.$$

稳健 t 统计量为

$$t_k = \frac{\hat{\beta}_{k,ols} - \beta_k^0}{SE^*(\hat{\beta}_{k,ols})},$$

其中，

$$SE^*(\hat{\beta}_{k,ols}) \equiv \sqrt{\frac{1}{n} \cdot \widehat{\mathrm{Avar}(\hat{\beta}_{k,ols})}} = \sqrt{s^2 \cdot \left((\mathbf{X}'\mathbf{X})^{-1}\right)_{kk}}.$$

于是，小样本 t 统计量与稳健 t 统计量相等，因此渐近服从标准正态分布。

类似地，

$$\begin{aligned}
W &= n \cdot (\mathbf{R}\hat{\boldsymbol{\beta}}_{ols} - \mathbf{r})' \left\{\mathbf{R}[\widehat{\mathrm{Avar}(\hat{\boldsymbol{\beta}}_{ols})}]\mathbf{R}'\right\}^{-1} (\mathbf{R}\hat{\boldsymbol{\beta}}_{ols} - \mathbf{r}) \\
&= n \cdot (\mathbf{R}\hat{\boldsymbol{\beta}}_{ols} - \mathbf{r})' \left\{\mathbf{R}[n \cdot s^2 \cdot (\mathbf{X}'\mathbf{X})^{-1}]\mathbf{R}'\right\}^{-1} (\mathbf{R}\hat{\boldsymbol{\beta}}_{ols} - \mathbf{r}) \\
&= (\mathbf{R}\hat{\boldsymbol{\beta}}_{ols} - \mathbf{r})' \left\{\mathbf{R}[(\mathbf{X}'\mathbf{X})^{-1}]\mathbf{R}'\right\}^{-1} (\mathbf{R}\hat{\boldsymbol{\beta}}_{ols} - \mathbf{r})/s^2 \\
&= J \cdot F \\
&= (RSS_R - RSS_U)/s^2.
\end{aligned}$$

于是，Wald 统计量 W 与 $J \cdot F$ 相等，因此，$J \cdot F$ 渐近服从 $\chi^2(J)$。 □

例 2.4 考虑例2.2中的解释变量的联合显著性检验。若条件同方差假设成立，式 (2.29) 已经证明：

$$F = \frac{R^2/(K-1)}{(1-R^2)/(n-K)}.$$

由上式，容易得到：

$$nR^2 = \frac{1}{\dfrac{n-K}{n} + \dfrac{1}{n}(K-1)F}(K-1)F.$$

由定理 2.14 中 (c) 的结论，$(K-1)F$ 依分布收敛，进而，$\frac{1}{n}(K-1)F \underset{p}{\to} 0$。[13] 于是，$nR^2$ 的渐近分布与 $(K-1)F$ 的渐近分布同为 $\chi^2(K-1)$。

（可行）广义最小二乘的大样本性质

为了讨论问题的方便，我们将假设 2.2* 和假设 2.5* 加强为 $\{y_i, \mathbf{x}_i\}$ 是 i.i.d. 的。我们考虑的条件二阶矩形式为式 (2.32)。

若 $\boldsymbol{\alpha}$ 已知，我们可以直接对数据进行转换得到如下回归方程：

$$\tilde{y}_i = \tilde{\mathbf{x}}_i'\boldsymbol{\beta} + \tilde{\varepsilon}_i, \tag{2.41}$$

其中，

$$\tilde{y}_i \equiv \frac{y_i}{\sqrt{\mathbf{z}_i'\boldsymbol{\alpha}}}, \quad \tilde{\mathbf{x}}_i \equiv \frac{\mathbf{x}_i}{\sqrt{\mathbf{z}_i'\boldsymbol{\alpha}}}, \quad \tilde{\varepsilon}_i \equiv \frac{\varepsilon_i}{\sqrt{\mathbf{z}_i'\boldsymbol{\alpha}}}.$$

[13] 见 Van der Vaart (2000) 中的定理 2.4。

对转换后的回归方程直接使用最小二乘得到如下加权最小二乘估计量：

$$\begin{aligned}\hat{\boldsymbol{\beta}}(\mathbf{V}) &\equiv \left(\frac{1}{n}\sum_{i=1}^{n}\tilde{\mathbf{x}}_i\tilde{\mathbf{x}}_i'\right)^{-1}\frac{1}{n}\sum_{i=1}^{n}\tilde{\mathbf{x}}_i\cdot\tilde{y}_i \\ &= \left(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y},\end{aligned}$$

其中，$\mathbf{V}=\operatorname{diag}(\mathbf{z}_1'\boldsymbol{\alpha},\ldots,\mathbf{z}_n'\boldsymbol{\alpha})$。如果将假设 2.3* 加强为：

$$\mathbb{E}(\varepsilon_i\mid\mathbf{x}_i)=0,$$

则有 $\mathbb{E}(\tilde{\varepsilon}_i\mid\tilde{\mathbf{x}}_i)=0$。若再假设 $\mathbb{E}(\tilde{\mathbf{x}}_i\tilde{\mathbf{x}}_i')$ 非奇异，我们有式 (2.41) 满足假设 2.1*-2.5*；同时，$\tilde{\varepsilon}_i$ 满足条件同方差而且 $\sigma^2=1$，即，$\mathbb{E}(\tilde{\varepsilon}_i^2\mid\tilde{\mathbf{x}}_i)=1$。所以，由定理 2.14，$\hat{\boldsymbol{\beta}}(\mathbf{V})$ 一致且渐近正态，并且其渐近方差为：

$$\begin{aligned}\operatorname{Avar}(\hat{\boldsymbol{\beta}}(\mathbf{V})) &= \sigma^2\mathbb{E}(\tilde{\mathbf{x}}_i\tilde{\mathbf{x}}_i')^{-1} \\ &= \mathbb{E}(\tilde{\mathbf{x}}_i\tilde{\mathbf{x}}_i')^{-1} \\ &= \operatorname{plim}\left(\frac{1}{n}\sum_{i=1}^{n}\tilde{\mathbf{x}}_i\tilde{\mathbf{x}}_i'\right)^{-1}=\operatorname{plim}\left(\frac{1}{n}\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\right)^{-1}.\end{aligned}$$

若 $\boldsymbol{\alpha}$ 未知，我们需要将 e_i^2 对 $\mathbf{z}_i$ 进行回归去估计 $\boldsymbol{\alpha}$，可以证明这样得到的估计量 $\hat{\boldsymbol{\alpha}}$ 是 $\boldsymbol{\alpha}$ 的一致估计量（见习题 7.）。$\boldsymbol{\beta}$ 的 FGLS 估计量可以表示为 $\hat{\boldsymbol{\beta}}(\hat{\mathbf{V}})$，其中，$\hat{\mathbf{V}}=\operatorname{diag}(\mathbf{z}_1'\hat{\boldsymbol{\alpha}},\ldots,\mathbf{z}_n'\hat{\boldsymbol{\alpha}})$。通过施加一些适当的额外假设（见 Amemiya (1977)），我们可以证明：

- $\sqrt{n}(\hat{\boldsymbol{\beta}}(\mathbf{V})-\boldsymbol{\beta})$ 和 $\sqrt{n}(\hat{\boldsymbol{\beta}}(\hat{\mathbf{V}})-\boldsymbol{\beta})$ 渐近等价，$\sqrt{n}(\hat{\boldsymbol{\beta}}(\hat{\mathbf{V}})-\boldsymbol{\beta})$ 的渐近分布与 $\sqrt{n}(\hat{\boldsymbol{\beta}}(\mathbf{V})-\boldsymbol{\beta})$ 的渐近分布相同；
- $\operatorname{plim}\left(\frac{1}{n}\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X}\right)^{-1}=\operatorname{plim}\left(\frac{1}{n}\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\right)^{-1}$，因此，$\left(\frac{1}{n}\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X}\right)^{-1}$ 是 $\operatorname{Avar}(\hat{\boldsymbol{\beta}}(\hat{\mathbf{V}}))$ 的一致估计量。

由定理 2.8，我们知道 GLS 比 OLS 有效。在大样本情况下，若不存在对误差项条件二阶矩的误设，我们可以证明 FGLS 比 OLS 渐近有效，即，FGLS 估计量的渐近方差小于或等于 OLS 估计量的渐近方差（见习题 8.）。

使用异方差稳健标准误或者使用 FGLS 都可以处理误差项存在条件异方差的情况，实证分析中我们到底应该使用哪一个是值得讨论的问题。FGLS 的优点是它是渐近有效的，前提是我们要正确设定误差项条件二阶矩的函数形式。如果条件二阶矩的函数形式设置有误，虽然 FGLS 估计量仍然是一致的，但是它的渐近方差可能大于 OLS 的渐近方差。此外，即使在大样本情况下，模型误设会导致 FGLS 估计量的标准误和检验统计量失效。使用异方差稳健标准误的优点是我们不需要知道条件异方差的具体函数形式，但是，与正确设定了条件异方差函数形式的 FGLS 估计量的标准误相比，异方差稳健标准误不是有效的。

正确设定条件异方差的函数形式在实证分析中几乎是不可能的，因此，现在通常的做法是直接使用异方差稳健标准误进行统计推断。这样做可能会损失效率，但是保证了统计推断结果的稳健性。如果核心解释变量的系数在使用异方差稳健标准误时是显著的，我们可以不去关心标准误的有效性，因为使用更有效的标准误只会让系数更加显著。但是，如果核心解释变量的系数在使用异方差稳健标准误时是不显著的，我们可以考虑尝

试使用 FGLS 进行修正。在条件异方差函数形式设置正确时，FGLS 估计的回归模型是满足条件同方差假设的；若担心条件异方差的函数形式误设，可以在使用 FGLS 的同时汇报 FGLS 回归的异方差稳健标准误。

2.7 总结

线性回归模型是计量经济学中最基础的模型，最小二乘估计法是估计线性回归模型系数的最常用方法。本章首先详细介绍了古典线性回归模型的假设，包括严格外生假设、无完全共线性假设和球形扰动假设等。这些假设保证了最小二乘估计量的无偏性和有效性。关于有效性，我们证明了著名的 Gauss-Markov 定理，该定理指出最小二乘估计量是所有线性无偏估计量中方差最小的估计量。著名计量学家 Bruce Hansen 最近发表在 *Econometrica* 上面的文章（见 Hansen (2022c)）提出了新的 Gauss-Markov 定理，该文章认为最小二乘估计量是所有无偏估计量中方差最小的估计量。此结论引起了一些学者的质疑，推荐感兴趣的读者阅读文献 Pötscher et al. (2022)。球形扰动假设包含条件同方差假设和无自相关假设，在实证分析时，条件异方差是普遍现象，金融时间序列的波动率聚集特征就是条件异方差的典型例子。此外，多重共线性是实证分析中可能遇到的数据问题，但是多重共线性是否会最终影响估计量的方差还取决于 σ^2 和解释变量样本的波动情况。因此，我们不应该单纯的依据解释变量的相关性决定对解释变量的取舍。

虽然严格外生性保证了最小二乘估计量的无偏性质，但是，对于时间序列数据来说，严格外生性一般是不成立的。在最小二乘估计量的大样本性质中，我们放松了严格外生性假设，取而代之的是前定假设。前定假设要求解释变量与同期的误差项正交，在此假设下，最小二乘估计量满足一致性。无偏性和一致性没有必然的联系，前者是小样本性质；后者是大样本性质。

本章的另外一个重要内容是多元线性回归模型的统计推断理论。在小样本下，我们通过施加正态假设推导出了单个系数假设检验的 t 统计量和一般线性假设检验的 F 统计量的精确分布。在大样本下，我们通过鞅差的中心极限定理建立基于异方差稳健一致协方差估计量的假设检验理论。我们推导了单个系数假设检验的稳健 t 统计量和一般线性假设检验的 Wald 统计量的渐近分布。同时，在同方差假设下，小样本 t 统计量与稳健 t 统计量相等；$J \cdot F$ 与 Wald 统计量相等。

正确计算标准误是进行统计推断的前提。在大样本统计推断中，我们使用了异方差稳健（或异方差一致）标准误。但是，由于鞅差假设，此标准误不适用于存在自相关的情况。这里的“自相关”使用的是时间序列的术语，实际上除了时间序列维度的自相关性，本章的异方差稳健标准误也不适用于存在横截面相关的情况。在计量模型中考虑横截面相关性是近年来理论计量研究的热点，与时间序列数据不同，横截面的相关性没有一个像时间一样的自然顺序。本书第 3 章会进一步讨论标准误的计算，我们将介绍异方差自相关稳健标准误和聚类标准误。

第二章习题

1. 如何解释“水平-对数”回归模型的系数？

2. 证明：若假设 2.3 成立，则 $\mathbf{X}'\mathbf{X}$ 是正定矩阵。
3. 证明：F 统计量可以表示为：
$$F = \frac{(R_U^2 - R_R^2)/J}{(1 - R_U^2)/(n-K)},$$
其中，R_U^2 和 R_R^2 分别表示无约束回归和有约束回归的 R^2。
4. 证明：式 (2.31) 满足假设 2.3。
5. 证明：式 (2.34) 是 $\{\mathbf{g}_i\}$ 为鞅差序列的充分条件。
6. 证明：式 (2.34) 可以推出误差项无自相关而且与同期和滞后期的解释变量也不存在相关性。
7. 证明：在 FGLS 中，回归 $e_i^2 = \mathbf{z}_i'\boldsymbol{\alpha} + \eta_i$ 的最小二乘估计量 $\hat{\boldsymbol{\alpha}}$ 是 $\boldsymbol{\alpha}$ 的一致估计量。
8. 证明：$\mathrm{Avar}(\hat{\boldsymbol{\beta}}(\hat{\mathbf{V}})) \preceq \mathrm{Avar}(\hat{\boldsymbol{\beta}}_{ols})$。

第 三 章 非球形扰动与标准误的计算

内容提要

- White 异方差检验 3.1
- Breusch-Pagan 异方差检验 3.2
- 修正的 Q 检验 3.4
- 异方差稳健标准误 3.3
- 异方差自相关稳健标准误 3.4
- 聚类标准误 3.5
- Jackknife 3.6
- Bootstrap 3.6

古典假设中的球形扰动假设要求误差项满足条件同方差和无条件自相关，对这两个条件是否满足的假设检验在计量经济学中分别称为异方差检验和自相关检验。本章第一节和第二节将分别介绍异方差检验和自相关检验的基本方法。球形扰动假设不成立不会影响最小二乘估计量的无偏性和一致性，但是，Gauss-Markov 定理不再成立（最小二乘估计量不是有效的），同时，基于此假设的小样本 t 检验和 F 检验失效。与有效性相比，统计推断往往更加关注结果的稳健性，本章第三节至第五节将分别讨论异方差稳健标准误、异方差自相关稳健标准误和聚类标准误的计算问题。此外，基于稳健标准误的统计推断通常使用检验统计量的渐近分布，在样本量有限的情况下，检验的真实显著性水平和功效（power）与名义值可能出现较大的偏离，本章最后讨论如何使用 Bootstrap 等重抽样方法进行标准误的计算。想进一步深入学习本章内容的读者可以阅读 Hansen (2022a) 和 Hayashi (2000)。

3.1 异方差检验

在第 2 章，我们介绍了基于异方差稳健标准误的大样本统计推断理论，稳健的 t 统计量和 Wald 统计量既适用于同方差也适用于异方差。因此，无论实际情况是同方差还是异方差，汇报和使用异方差稳健标准误已经成为实证分析的常规做法。然而，在异方差条件下，最小二乘估计量不是最有效的，这时可以考虑采用 FGLS。因此，对异方差进行检验依然具有一定的现实意义。

本节将要介绍两种常用的异方差检验：White 异方差检验和 Breusch-Pagan 异方差检验。

White 异方差检验

由定理 2.10，$\hat{\mathbf{S}} = \frac{1}{n}\sum_{i=1}^{n} e_i^2\mathbf{x}_i\mathbf{x}_i'$ 是 $\mathbf{S} = \mathbb{E}(\varepsilon_i^2\mathbf{x}_i\mathbf{x}_i')$ 的一致估计。另一方面，在同方差假设 2.7* 成立的情况下，$\mathbf{S} = \sigma^2\boldsymbol{\Sigma}_{\mathbf{xx}}$，其一致估计为 $s^2\mathbf{S}_{\mathbf{xx}}$。若条件同方差的原假设成立，则：

$$\hat{\mathbf{S}} - s^2\mathbf{S}_{\mathbf{xx}} = \frac{1}{n}\sum_{i=1}^{n}(e_i^2 - s^2)\mathbf{x}_i\mathbf{x}_i' \underset{p}{\to} \mathbf{0}.$$

White 异方差检验（见 White (1980)）的基本思想是比较 $\hat{\mathbf{S}}$ 和 $s^2\mathbf{S}_{\mathbf{xx}}$ 之间的马氏距离，若此距离足够大，则拒绝同方差的原假设。具体地，将 $K \times K$ 维对称阵 $\mathbf{x}_i\mathbf{x}_i'$ 中彼此不同的非常数元素组成一个列向量，记为 $\boldsymbol{\psi}_i$。定义：

$$\mathbf{c}_n \equiv \frac{1}{n}\sum_{i=1}^{n}(e_i^2 - s^2)\boldsymbol{\psi}_i,$$

若条件同方差的原假设成立，我们有 $\mathbf{c}_n \underset{p}{\to} \mathbf{0}$。在一些适当的假设条件下，$\sqrt{n}\mathbf{c}_n$ 依分布收敛到正态分布，期望是 $\mathbf{0}$，渐近方差为某一正定矩阵 $\mathbf{B}$。因此，

$$n \cdot \mathbf{c}_n'\hat{\mathbf{B}}^{-1}\mathbf{c}_n \underset{d}{\to} \chi^2(m),$$

其中，$\hat{\mathbf{B}}$ 是 $\mathbf{B}$ 的一致估计，m 是向量 $\boldsymbol{\psi}_i$ 的维度。若恰当选取 $\hat{\mathbf{B}}$，$n \cdot \mathbf{c}_n'\hat{\mathbf{B}}^{-1}\mathbf{c}_n$ 可以通过 e_i^2 对常数项和 $\boldsymbol{\psi}_i$ 的回归的 nR^2 得到。

定理 3.1. White 异方差检验

令假设 2.1* 和 2.4* 成立，同时，假设：（1）$\{y_i, \mathbf{x}_i\}$ 满足 i.i.d.，且 $\mathbb{E}(\varepsilon_i^2\mathbf{x}_i\mathbf{x}_i')$ 有限；（2）ε_i 独立于 $\mathbf{x}_i$；（3）ε_i 和 $\mathbf{x}_i$ 满足一定的矩条件[a]，则有：

$$nR^2 \underset{d}{\to} \chi^2(m),$$

其中，R^2 来自 e_i^2 对常数项和 $\boldsymbol{\psi}_i$ 的回归；m 是向量 $\boldsymbol{\psi}_i$ 的维度。

[a] 具体条件请参考 White (1980)。

♡

注 在 Stata 中，我们可以在回归后使用命令 “estat imtest, white” 得到 White 异方差检验的结果。

Breusch-Pagan 异方差检验

考虑如下模型设定：

$$y_i = \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i, \tag{3.1}$$

其中，$\{\varepsilon_i\}_{i=1}^n$ 服从零均值的正态分布且彼此独立，ε_i 的方差为：

$$\sigma_i^2 = \sigma^2 \cdot h(\mathbf{z}_i'\boldsymbol{\alpha}), \tag{3.2}$$

其中，$\boldsymbol{\alpha}$ 是 $p \times 1$ 维的向量，$\mathbf{z}_i$ 的第一个元素为 1，$\mathbf{x}_i$ 和 $\mathbf{z}_i$ 都为外生变量[1]。条件同方差的原假设可以表示为：

$$\mathbb{H}_0 : \alpha_2 = \cdots = \alpha_p = 0.$$

这是因为当 $\mathbb{H}_0$ 成立时，$\sigma_i^2 = \sigma^2 h(\alpha_1)$ 为常数。Breusch et al. (1979) 构造了检验 $\mathbb{H}_0$ 的检验统计量并给出了它的渐近分布。

[1] $\mathbf{z}_i$ 中的元素形式比较灵活，可以是 $\mathbf{x}_i$ 中的某些变量，也可以是 $\mathbf{x}_i$ 中变量的交乘。

定理 3.2. Breusch-Pagan 异方差检验

考虑式 (3.1) 和 (3.2) 的模型设定。令 $s^2 = \boldsymbol{e}'\boldsymbol{e}/n$，$g_i = e_i^2/s^2$。检验 $\mathbb{H}_0 : \alpha_2 = \cdots = \alpha_p = 0$ 的拉格朗日乘数统计量[a]为：

$$\mathrm{LM} \equiv \frac{1}{2}\left[\boldsymbol{g}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\boldsymbol{g} - n\right],$$

其中，$\boldsymbol{g} = (g_1, \ldots, g_n)'$，$\mathbf{Z} = (\mathbf{z}_1', \ldots, \mathbf{z}_n')'$。在 $\mathbb{H}_0$ 成立的情况下，LM 渐近服从 $\chi^2(p-1)$ 分布。此外，$\boldsymbol{g}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\boldsymbol{g} - n$ 等于 g_i 对 $\mathbf{z}_i$ 回归的回归平方和。

[a] 见本书第7章7.4节。

♡

注 在 Stata 中，我们可以在回归后使用如下命令实现 Breusch-Pagan 异方差检验：

- estat hettest：使用拟合值 $\hat{y}_i$ 作为 $\mathbf{z}_i$；
- estat hettest, rhs：使用解释变量 $\mathbf{x}_i$ 作为 $\mathbf{z}_i$；
- estat hettest varlist：使用用户设置的 varlist 中的变量作为 $\mathbf{z}_i$。

Breusch-Pagan 异方差检验只适用于误差项服从正态分布的情况，Koenker (1981) 将正态假设放松为误差项独立同分布且具有有限四阶矩，此改进的检验可以通过“estat hettest, iid”实现。此外，“estat hettest, fstat”实现了 Wooldridge (2019) 中第 270 页的 F 统计量，其渐近分布的推导也不需要正态假设。

事实上，White 异方差检验可以看成是放松了正态假设的 Breusch-Pagan 异方差检验的特例。在 White 异方差检验中，$\mathbf{z}_i$ 被设置成解释变量、解释变量的平方以及解释变量之间的交乘。

例 3.1 考虑 Wooldridge (2010) 中例 4.1 使用的数据集 mroz.dta，我们将工资对数（lwage）对受教育年限（educ）和工作经验（exper）做回归，并基于回归的结果进行异方差检验，使用的 Stata 程序和输出结果如下：

```
. use mroz.dta, replace
. keep if inlf == 1 //只保留有工作的女性样本
. quietly reg lwage educ exper
. estat hettest //Breusch-Pagan异方差检验

Breusch-Pagan/Cook-Weisberg test for heteroskedasticity
Assumption: Normal error terms
Variable: Fitted values of lwage

H0: Constant variance

    chi2(1) =  2.05
Prob > chi2 = 0.1525

. estat hettest, iid //放松正态假设的Breusch-Pagan异方差检验
```

```

Breusch-Pagan/Cook-Weisberg test for heteroskedasticity
Assumption: i.i.d. error terms
Variable: Fitted values of lwage

H0: Constant variance

    chi2(1) =  0.71
Prob > chi2 = 0.3986

. estat imtest, white //White异方差检验

White's test
H0: Homoskedasticity
Ha: Unrestricted heteroskedasticity

    chi2(5) = 13.12
Prob > chi2 = 0.0223

. estat hettest educ exper c.educ#c.educ c.exper#c.exper ///
               c.educ#c.exper, iid //用hettest命令实现White异方差检验

Breusch-Pagan/Cook-Weisberg test for heteroskedasticity
Assumption: i.i.d. error terms
Variables: educ exper c.educ#c.educ c.exper#c.exper c.educ#c.exper

H0: Constant variance

    chi2(5) = 13.12
Prob > chi2 = 0.0223
```

3.2 自相关检验

上一章指出，若回归中有常数项，假设 2.5* 成立意味着误差项不存在自相关。自相关检验是检验误差项是否存在自相关的假设检验，若拒绝误差项不存在自相关的原假设，则假设 2.5* 不成立。因为自相关是时间序列分析使用的术语，所以我们将区分不同观测的下标 i 改为 t，样本量为 T。

可观测时间序列的自相关检验

我们首先讨论可观测时间序列的自相关检验。假设标量（scalar）弱平稳随机过程的一个样本为 $\{r_t\}_{t=1}^T$，r_t 的滞后 l 期的样本自协方差（lag-l sample autocovariance）为：

$$\hat{\gamma}_l \equiv \frac{1}{T}\sum_{t=l+1}^{T}(r_t-\bar{r})(r_{t-l}-\bar{r}), \quad 0 \leqslant l < T-1,$$

其中，

$$\bar{r} \equiv \frac{1}{T}\sum_{t=1}^{T} r_t.$$

r_t 的滞后 l 期的样本自相关系数（lag-l sample autocorrelation）为：

$$\hat{\rho}_l \equiv \frac{\hat{\gamma}_l}{\hat{\gamma}_0}.$$

如果 $\{r_t\}$ 是来自遍历平稳的随机过程的样本，则 $\hat{\gamma}_l \underset{p}{\to} \gamma_l$，$\hat{\rho}_l \underset{p}{\to} \rho_l$，其中，$\gamma_l \equiv \mathrm{Cov}(r_t, r_{t-l})$ 是滞后 l 期的自协方差（lag-l autocovariance），$\rho_l \equiv \gamma_l/\gamma_0$ 是滞后 l 期的自相关系数（lag-l autocorrelation）。

定理 3.3. $\hat{\gamma}_l$ 和 $\hat{\rho}_l$ 的中心极限定理

假设 $\{r_t\}$ 可以表示为 $\mu+\varepsilon_t$，其中，ε_t 是遍历[a]平稳的鞅差序列并且满足：

$$\mathbb{E}(\varepsilon_t^2 \mid \varepsilon_{t-1}, \varepsilon_{t-2}, \ldots) = \sigma^2 > 0,$$

则有，

$$\sqrt{T}\hat{\boldsymbol{\gamma}} \underset{d}{\to} \mathcal{N}(\mathbf{0}, \sigma^4\mathbf{I}_p) \quad 和 \quad \sqrt{T}\hat{\boldsymbol{\rho}} \underset{d}{\to} \mathcal{N}(\mathbf{0}, \mathbf{I}_p),$$

其中，$\hat{\boldsymbol{\gamma}} = (\hat{\gamma}_1, \ldots, \hat{\gamma}_p)'$，$\hat{\boldsymbol{\rho}} = (\hat{\rho}_1, \ldots, \hat{\rho}_p)'$。

[a] 事实上，保证此定理结论成立并不需要遍历条件，参见 Hall et al. (1980) 的定理 6.7。

自相关检验的一种方式是检验单个自相关系数是否等于 0，即，$\mathbb{H}_0: \rho_k = 0$。由定理 3.3，我们有：

$$\sqrt{T}\hat{\rho}_k \underset{d}{\to} \mathcal{N}(0,1).$$

绘制样本自相关系数函数（autocorrelation function, ACF）图是汇报单个自相关系数检验结果的常用方式。样本 ACF 图的横轴为滞后期，纵轴为每个滞后期对应的样本自相关系数，通过图中给出的置信域可以直观地判断是拒绝还是不能拒绝原假设。

例 3.2 我们用上证指数从 2007 年 4 月 13 日至 2019 年 12 月 31 日的周对数收益率为例说明如何利用 Stata 和 R 绘制样本 ACF 图，使用的程序如下：

```
//Stata程序
. import delimited "上证指数收益率.csv", clear
. g time = _n        //生成时间变量
. tsset time         //设置时间变量
. ac rtn, lags(10) //绘制滞后1-10期的样本ACF图
```

```r
#R程序
shindex <- read.table("~/上证指数收益率.csv",sep = ",",header = TRUE)
library(forecast)
Return <- shindex$rtn
ggAcf(Return,10)
```

两种语言绘制的样本 ACF 图如图 3.1。左图为 Stata 绘制的样本 ACF 图，灰色阴影部分为 95% 的置信域；右图为 R 绘制的样本 ACF 图，上下两条蓝色虚线高度分别为 $1.96/\sqrt{T}$ 和 $-1.96/\sqrt{T}$，也就是基于标准正态分布构造的 95% 的置信域的上下界。注意，与右图不同，左图阴影区域的上下两条边界并非直线，这是由于 Stata 在构造置信域时使用的是 Bartlett 公式，对于给定的滞后阶数 k，置信区间的上下界为 $\pm 1.96 \cdot \sqrt{(1+2\sum_{i=1}^{k-1}\hat{\rho}_i^2)/T}$，关于 Bartlett 公式的详细介绍请参考 Brockwell et al. (1991) 的第 7.2 节。

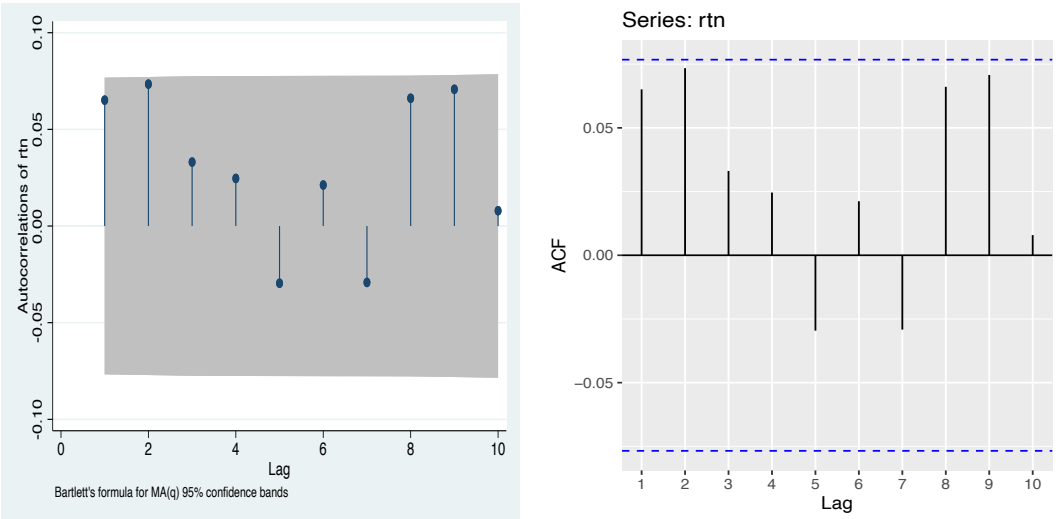

图 3.1: 上证指数收益率样本 ACF 图（左图：Stata；右图：R）

另一种更常用的自相关检验是检验多个自相关系数是否同时为 0，检验的原假设为 $\mathbb{H}_0: \rho_1 = \rho_2 = \cdots = \rho_p = 0$。由定理 3.3，向量 $\sqrt{T}\hat{\boldsymbol{\rho}}$ 的每个元素都渐近服从标准正态分布且彼此之间渐近独立。因此，我们有：

$$Q_*(p) \equiv T\sum_{l=1}^{p}\hat{\rho}_l^2 = \sum_{l=1}^{p}(\sqrt{T}\hat{\rho}_l)^2 \xrightarrow[d]{} \chi^2(p).$$

$Q_*(p)$ 被称为 Box-Pierce Q 检验统计量（见 Box et al. (1970)）。Ljung et al. (1978) 对 $Q_*(p)$ 进行了改进，提出了如下的 Ljung-Box Q 检验统计量：

$$Q(p) \equiv T(T+2)\sum_{l=1}^{p}\frac{\hat{\rho}_l^2}{T-l} = \sum_{l=1}^{p}\frac{T+2}{T-l}(\sqrt{T}\hat{\rho}_l)^2 \xrightarrow[d]{} \chi^2(p).$$

通常，在样本有限的情况下，Ljung-Box Q 检验比 Box-Pierce Q 检验具有更高的检验功效。两种 Q 检验也被称为混合（portmanteau）检验。

例 3.3 我们用上证指数从 2007 年 4 月 13 日至 2019 年 12 月 31 日的周对数收益率为例说明如何利用 Stata 和 R 进行混合检验，使用的程序如下：

```
//接着例3.2的Stata程序
. wntestq rtn      //Ljung-Box检验

Portmanteau (Q) statistic = 57.3946
Prob > chi2(40)          =     0.0367
```

```
#接着例3.2的R程序
Box.test(Return,lag=40,type='Ljung-Box')
#输出的p值为0.0367，与Stata程序结果相同
```

注 混合检验滞后阶数的选择对检验的结果有一定的影响，若阶数选择过小，可能无法检验到更高阶的相关性；若阶数过大，检验统计量在小样本下的实际分布可能严重偏离卡方分布。Stata 的 wntestq 命令默认的阶数为 $\min(\lfloor n/2\rfloor-2,40)$，其中，$\lfloor n/2\rfloor$ 表示小于或等于 $n/2$ 的最大整数。

回归误差项的自相关检验

对于假设 2.1*-2.5* 考虑的线性回归模型，如果误差项 ε_t 可观测，我们可以计算：

$$\tilde{\rho}_l \equiv \frac{\tilde{\gamma}_l}{\tilde{\gamma}_0}, \quad l=1,2,\ldots,$$

其中，

$$\tilde{\gamma}_l \equiv \frac{1}{T}\sum_{t=l+1}^{T}\varepsilon_t\varepsilon_{t-l}, \quad l=0,1,2,\ldots. \quad \text{（注意误差项的均值为 0）}$$

然而，由于误差项无法观测，可以考虑用残差 e_t 替换 ε_t 并计算样本自相关系数：

$$\hat{\rho}_l \equiv \frac{\hat{\gamma}_l}{\hat{\gamma}_0}, \quad l=1,2,\ldots,$$

其中，

$$\hat{\gamma}_l \equiv \frac{1}{T}\sum_{t=l+1}^{T}e_te_{t-l}, \quad l=0,1,2,\ldots. \quad \text{（注意残差项的样本均值为 0）}$$

直观上，我们可以基于 $\hat{\rho}_l$ 进行 Ljung-Box 混合检验去推断误差项是否存在自相关。但是，由于残差是利用最小二乘估计量计算得到的，为了保证 Ljung-Box 检验统计量的渐近分布依然有效，我们需要严格外生条件。事实上，因为

$$e_t=\varepsilon_t-\mathbf{x}_t'(\hat{\boldsymbol{\beta}}_{ols}-\boldsymbol{\beta}),$$

所以，我们有：

$$\begin{aligned}\hat{\gamma}_l &= \frac{1}{T}\sum_{t=l+1}^{T}e_te_{t-l}=\frac{1}{T}\sum_{t=l+1}^{T}\left[\varepsilon_t-\mathbf{x}_t'(\hat{\boldsymbol{\beta}}_{ols}-\boldsymbol{\beta})\right]\left[\varepsilon_{t-l}-\mathbf{x}_{t-l}'(\hat{\boldsymbol{\beta}}_{ols}-\boldsymbol{\beta})\right]\\ &= \tilde{\gamma}_l-\frac{1}{T}\sum_{t=l+1}^{T}(\mathbf{x}_{t-l}\varepsilon_t+\mathbf{x}_t\varepsilon_{t-l})'(\hat{\boldsymbol{\beta}}_{ols}-\boldsymbol{\beta})\end{aligned}$$

$$+ (\hat{\boldsymbol{\beta}}_{ols} - \boldsymbol{\beta})' \left(\frac{1}{T} \sum_{t=l+1}^{T} \mathbf{x}_t \mathbf{x}_{t-l}' \right) (\hat{\boldsymbol{\beta}}_{ols} - \boldsymbol{\beta}).$$

注意到 $\hat{\boldsymbol{\beta}}_{ols} - \boldsymbol{\beta} \underset{p}{\to} \mathbf{0}$，因此，$\hat{\gamma}_l - \tilde{\gamma}_l \underset{p}{\to} 0$，$(j = 0, 1, 2, \ldots)$。然而，

$$\begin{aligned} \sqrt{T}\hat{\gamma}_l &= \sqrt{T}\tilde{\gamma}_l - \frac{1}{T} \sum_{t=l+1}^{T} (\mathbf{x}_{t-l}\varepsilon_t + \mathbf{x}_t\varepsilon_{t-l})' \sqrt{T}(\hat{\boldsymbol{\beta}}_{ols} - \boldsymbol{\beta}) \\ &\quad + \underbrace{\sqrt{T}(\hat{\boldsymbol{\beta}}_{ols} - \boldsymbol{\beta})' \left(\frac{1}{T} \sum_{t=l+1}^{T} \mathbf{x}_t \mathbf{x}_{t-l}' \right) (\hat{\boldsymbol{\beta}}_{ols} - \boldsymbol{\beta})}_{\underset{p}{\to} 0}. \end{aligned}$$

注意到 $\sqrt{T}(\hat{\boldsymbol{\beta}}_{ols} - \boldsymbol{\beta})$ 服从渐近正态，因此依概率有界。为使 $\sqrt{T}\hat{\gamma}_l$ 和 $\sqrt{T}\tilde{\gamma}_l$ 之间的差别可忽略不计，我们需要 $(1/T)\sum_{t=l+1}^{T}(\mathbf{x}_{t-l}\varepsilon_t + \mathbf{x}_t\varepsilon_{t-l})$ 依概率收敛到 $\mathbf{0}$。若解释变量满足如下意义的严格外生性：对任意的 $t, s \in \{1, \ldots, T\}$，$\mathbb{E}(\mathbf{x}_t\varepsilon_s) = \mathbf{0}$，则有：

$$\frac{1}{T} \sum_{t=l+1}^{T} (\mathbf{x}_{t-l}\varepsilon_t + \mathbf{x}_t\varepsilon_{t-l}) \underset{p}{\to} \mathbb{E}(\mathbf{x}_{t-l}\varepsilon_t) + \mathbb{E}(\mathbf{x}_t\varepsilon_{t-l}) = \mathbf{0}.$$

于是，在严格外生条件下，我们依然可以使用 Ljung-Box Q 检验统计量及其渐近分布进行自相关检验。但是，若严格外生性不成立，基于残差的 Q 检验统计量可能不服从渐近卡方分布。

鉴于时间序列数据一般不满足严格外生性，我们需要对 Q 检验统计量进行修正。为此，我们需要施加如下假设：

假设 2.8*：（强前定和强条件同方差）

强前定：

$$\mathbb{E}(\varepsilon_t \mid \varepsilon_{t-1}, \varepsilon_{t-2}, \ldots, \mathbf{x}_t, \mathbf{x}_{t-1}, \ldots) = 0. \tag{3.3}$$

强条件同方差：

$$\mathbb{E}(\varepsilon_t^2 \mid \varepsilon_{t-1}, \varepsilon_{t-2}, \ldots, \mathbf{x}_t, \mathbf{x}_{t-1}, \ldots) = \sigma^2 > 0. \tag{3.4}$$

注 条件 (3.3) 强于假设 2.3* 和 2.5*；条件 (3.4) 强于假设 2.7*。

定理 3.4. 修正的 Q 检验

若假设 2.1*、2.2*、2.4* 和 2.8* 成立，则基于残差 e_t 计算的 $\hat{\gamma}_l$ 和 $\hat{\rho}_l$ 满足：

$$\sqrt{T}\hat{\boldsymbol{\gamma}} \underset{d}{\to} \mathcal{N}(\mathbf{0}, \sigma^4 \cdot (\mathbf{I}_p - \boldsymbol{\Phi})) \quad 和 \quad \sqrt{T}\hat{\boldsymbol{\rho}} \underset{d}{\to} \mathcal{N}(\mathbf{0}, \mathbf{I}_p - \boldsymbol{\Phi}),$$

其中，$\hat{\boldsymbol{\gamma}} = (\hat{\gamma}_1, \ldots, \hat{\gamma}_p)'$，$\hat{\boldsymbol{\rho}} = (\hat{\rho}_1, \ldots, \hat{\rho}_p)'$，$p \times p$ 维矩阵 $\boldsymbol{\Phi}$ 的第 (j, k) 位置上的元素 ϕ_{jk} 为：

$$\phi_{jk} = \mathbb{E}(\mathbf{x}_t\varepsilon_{t-j})'\mathbb{E}(\mathbf{x}_t\mathbf{x}_t')^{-1}\mathbb{E}(\mathbf{x}_t\varepsilon_{t-k})/\sigma^2.$$

♡

由遍历定理 B.7，如下定义的 $\widehat{\boldsymbol{\Phi}}$ 可以一致估计 $\boldsymbol{\Phi}$：

$$\widehat{\boldsymbol{\Phi}} \equiv (\hat{\phi}_{jk}), \quad \hat{\phi}_{jk} \equiv \bar{\boldsymbol{\mu}}_j' \mathbf{S}_{\mathbf{xx}}^{-1} \bar{\boldsymbol{\mu}}_k / s^2, \quad (j, k = 1, 2, \ldots, p),$$

其中，

$$s^2 \equiv \frac{1}{T-K}\sum_{t=1}^{T} e_t^2, \quad \bar{\boldsymbol{\mu}}_j \equiv \frac{1}{T}\sum_{t=j+1}^{T} \mathbf{x}_t \cdot e_{t-j}.$$

于是，我们容易构造如下修正的 Q 检验统计量并得到其渐近分布：

$$\text{修正的 Box-Pierce } Q \equiv T \cdot \hat{\boldsymbol{\rho}}'(\mathbf{I}_p - \widehat{\boldsymbol{\Phi}})^{-1}\hat{\boldsymbol{\rho}} \underset{d}{\to} \chi^2(p).$$

修正的 Box-Pierce Q 检验统计量也可以通过回归的方式得到，考虑如下辅助回归：

$$e_t = \mathbf{x}_t'\boldsymbol{\phi} + \theta_1 e_{t-1} + \cdots + \theta_p e_{t-p} + \nu_t, \quad t = p+1, \ldots, T. \tag{3.5}$$

在定理 3.4 的假设条件下，修正的 Box-Pierce Q 检验统计量与 $p \cdot F$ 渐近等价，其中，F 为检验辅助回归 (3.5) 中 $e_{t-1}, \ldots, e_{t-p}$ 系数联合显著性的 F 统计量。此外，$p \cdot F$ 与辅助回归 (3.5) 的 $T \cdot R^2$ 渐近等价，基于 $T \cdot R^2$ 的检验称为 Breusch-Godfrey 自相关检验（见 Breusch (1978) 和 Godfrey (1978)）。

现证明 $p \cdot F$ 与辅助回归 (3.5) 的 $T \cdot R^2$ 渐近等价。注意到：

$$p \cdot F = \frac{RSS_R - RSS_U}{RSS_U/(T-K-2p)} = (T-K-2p)\frac{RSS_R - RSS_U}{RSS_U},$$

其中，RSS_U 由回归 (3.5) 得到，RSS_R 由 e_t 仅对 $\mathbf{x}_t$ 进行回归的约束回归得到。由于残差 e_t 正交于 $\mathbf{x}_t$，所以 $RSS_R = \boldsymbol{e}'\boldsymbol{e}$，并且辅助回归 (3.5) 的非中心化 R^2 为：

$$R_{uc}^2 = \frac{\boldsymbol{e}'\boldsymbol{e} - RSS_U}{\boldsymbol{e}'\boldsymbol{e}}.$$

于是，我们有：

$$p \cdot F = (T-K-2p)\frac{R_{uc}^2}{1-R_{uc}^2}.$$

又因为 e_t 的样本均值为 0，R_{uc}^2 等于 R^2，则有：

$$p \cdot F = (T-K-2p)\frac{R^2}{1-R^2}.$$

上式经整理可得：

$$T \cdot R^2 = \frac{T}{T-K-2p} \cdot \frac{1}{1+\dfrac{p \cdot F}{T-K-2p}} \cdot p \cdot F.$$

由上式容易知道 $p \cdot F - T \cdot R^2 \underset{p}{\to} 0$。

例 3.4 我们用上证指数从 2007 年 4 月 13 日至 2019 年 12 月 31 日的周对数收益率对其滞后一期做回归并检验回归的误差项是否存在自相关，Stata 和 R 的程序分别如下：

```
//接着例3.2的Stata程序
. gen rtn_lag = L.rtn        //生成rtn的滞后一期
. qui reg rtn rtn_lag        //一阶自回归
. estat bgodfrey, lags(10) //Breusch-Godfrey自相关检验，p为10
//输出的p值为0.3596
```

```
#R程序
shindex <- read.table("~/上证指数收益率.csv",sep = ",",header = TRUE)
```

```
date <- as.Date(as.character(shindex$date),format="%Y/%m/%d")
rtn <- xts::xts(shindex$rtn,date) #把rtn转换为xts对象
rtn_lag <- lag(rtn,1)             #将rtn滞后一期
mydata <- as.data.frame(rtn,rtn_lag)
lmtest::bgtest(rtn~rtn_lag, order=10, data=mydata)
#输出的p值也为0.3596
```

3.3 异方差稳健标准误

在例 3.1 中，我们将工资对数对受教育年限和工作经验做回归，并基于回归的结果进行了 Breusch-Pagan 异方差检验和 White 异方差检验。我们发现两个检验得到的结论是不一致的：在 5% 的显著性水平下，White 检验拒绝原假设，而 Breusch-Pagan 检验不能拒绝原假设。为了统计推断结论的稳健，现代实证分析的通常做法是直接使用异方差稳健标准误。

由定理 2.9，最小二乘估计量的渐近方差为：

$$\widehat{\mathrm{Avar}(\hat{\boldsymbol{\beta}}_{ols})} = \mathbf{S}_{\mathbf{xx}}^{-1}\hat{\mathbf{S}}\mathbf{S}_{\mathbf{xx}}^{-1} = \left(\frac{1}{n}\mathbf{X}'\mathbf{X}\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} e_i^2\mathbf{x}_i\mathbf{x}_i'\right)\left(\frac{1}{n}\mathbf{X}'\mathbf{X}\right)^{-1}.$$

因此，最小二乘估计量的异方差稳健协方差矩阵估计量为：

$$\widehat{\mathbf{V}}_{\hat{\boldsymbol{\beta}}_{ols}}^{\mathrm{HC0}} = (\mathbf{X}'\mathbf{X})^{-1}\left(\sum_{i=1}^{n} e_i^2\mathbf{x}_i\mathbf{x}_i'\right)(\mathbf{X}'\mathbf{X})^{-1}.$$

式 (2.39) 的 White 标准误是 $\widehat{\mathbf{V}}_{\hat{\boldsymbol{\beta}}_{ols}}^{\mathrm{HC0}}$ 的对角线上第 k 个元素的平方根。

由本章附录 2 中的式 (3.43)，e_i^2 是 σ^2 的有偏估计量，因此，残差平方的样本均值 $\hat{\sigma}^2$ 也是 σ^2 的有偏估计量（见式 (3.44)）。为了得到 σ^2 的无偏估计，我们需要在 $\hat{\sigma}^2$ 的基础上乘以 $n/(n-K)$ 进行调整。对 $\widehat{\mathbf{V}}_{\hat{\boldsymbol{\beta}}_{ols}}^{\mathrm{HC0}}$ 做同样的调整可以得到：

$$\widehat{\mathbf{V}}_{\hat{\boldsymbol{\beta}}_{ols}}^{\mathrm{HC1}} = \left(\frac{n}{n-K}\right)(\mathbf{X}'\mathbf{X})^{-1}\left(\sum_{i=1}^{n} e_i^2\mathbf{x}_i\mathbf{x}_i'\right)(\mathbf{X}'\mathbf{X})^{-1}. \tag{3.6}$$

利用附录 2 中定义的标准化残差和“去一”残差，我们还可以得到另外两种异方差稳健协方差矩阵估计量：

$$\widehat{\mathbf{V}}_{\hat{\boldsymbol{\beta}}_{ols}}^{\mathrm{HC2}} = (\mathbf{X}'\mathbf{X})^{-1}\left(\sum_{i=1}^{n} \bar{e}_i^2\mathbf{x}_i\mathbf{x}_i'\right)(\mathbf{X}'\mathbf{X})^{-1}, \tag{3.7}$$

和

$$\widehat{\mathbf{V}}_{\hat{\boldsymbol{\beta}}_{ols}}^{\mathrm{HC3}} = (\mathbf{X}'\mathbf{X})^{-1}\left(\sum_{i=1}^{n} \tilde{e}_i^2\mathbf{x}_i\mathbf{x}_i'\right)(\mathbf{X}'\mathbf{X})^{-1}. \tag{3.8}$$

$\widehat{\mathbf{V}}_{\hat{\boldsymbol{\beta}}_{ols}}^{\mathrm{HC1}}$ 对 $\widehat{\mathbf{V}}_{\hat{\boldsymbol{\beta}}_{ols}}^{\mathrm{HC0}}$ 进行的自由度调整是由 Hinkley (1977) 建议使用的，Stata 默认汇报的异方差稳健标准误即为 $\widehat{\mathbf{V}}_{\hat{\boldsymbol{\beta}}_{ols}}^{\mathrm{HC1}}$ 的对角线上元素的平方根，可以通过在回归命令 reg 后加“, r”实现；$\widehat{\mathbf{V}}_{\hat{\boldsymbol{\beta}}_{ols}}^{\mathrm{HC2}}$ 由 Horn et al. (1975) 提出，基于它的标准误在 Stata 中可以通过在回归命令 reg 后加“, vce (hc2)”实现；$\widehat{\mathbf{V}}_{\hat{\boldsymbol{\beta}}_{ols}}^{\mathrm{HC3}}$ 由 MacKinnon et al. (1985) 提出，基于它的标准误在

Stata 中可以通过在回归命令 reg 后加 “, vce (hc3)” 实现。

由附录 2 中给出的杠杆值性质，$0 < h_{ii} \leqslant 1$，[2] 因此，在 h_{ii} 不取极端值 1 的情况下，$(1-h_{ii})^{-2} > (1-h_{ii})^{-1} > 1$。容易验证：

$$\widehat{\mathbf{V}}^{\mathrm{HC0}}_{\hat{\boldsymbol{\beta}}_{ols}} \prec \widehat{\mathbf{V}}^{\mathrm{HC2}}_{\hat{\boldsymbol{\beta}}_{ols}} \prec \widehat{\mathbf{V}}^{\mathrm{HC3}}_{\hat{\boldsymbol{\beta}}_{ols}}. \tag{3.9}$$

在球形扰动假设下，我们有：

$$\begin{aligned}
\mathbb{E}\left[\widehat{\mathbf{V}}^{\mathrm{HC0}}_{\hat{\boldsymbol{\beta}}_{ols}} \mid \mathbf{X}\right] &= \left(\mathbf{X}'\mathbf{X}\right)^{-1}\left(\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i'\mathbb{E}[e_i^2 \mid \mathbf{X}]\right)\left(\mathbf{X}'\mathbf{X}\right)^{-1} \\
&= \left(\mathbf{X}'\mathbf{X}\right)^{-1}\left(\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i'(1-h_{ii})\sigma^2\right)\left(\mathbf{X}'\mathbf{X}\right)^{-1} \qquad \text{(由式 (3.43))} \\
&= \sigma^2\left(\mathbf{X}'\mathbf{X}\right)^{-1} - \sigma^2\left(\mathbf{X}'\mathbf{X}\right)^{-1}\left(\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i'h_{ii}\right)\left(\mathbf{X}'\mathbf{X}\right)^{-1} \\
&\prec \sigma^2\left(\mathbf{X}'\mathbf{X}\right)^{-1} = \mathrm{Var}(\hat{\boldsymbol{\beta}}_{ols} \mid \mathbf{X}).
\end{aligned}$$

类似地，我们可以证明：在球形扰动假设下，

$$\mathbb{E}\left[\widehat{\mathbf{V}}^{\mathrm{HC2}}_{\hat{\boldsymbol{\beta}}_{ols}} \mid \mathbf{X}\right] = \sigma^2\left(\mathbf{X}'\mathbf{X}\right)^{-1} = \mathrm{Var}(\hat{\boldsymbol{\beta}}_{ols} \mid \mathbf{X}), \tag{3.10}$$

和

$$\mathbb{E}\left[\widehat{\mathbf{V}}^{\mathrm{HC3}}_{\hat{\boldsymbol{\beta}}_{ols}} \mid \mathbf{X}\right] \succeq \sigma^2\left(\mathbf{X}'\mathbf{X}\right)^{-1} = \mathrm{Var}(\hat{\boldsymbol{\beta}}_{ols} \mid \mathbf{X}). \tag{3.11}$$

通常，$\widehat{\mathbf{V}}^{\mathrm{HC1}}_{\hat{\boldsymbol{\beta}}_{ols}}$、$\widehat{\mathbf{V}}^{\mathrm{HC2}}_{\hat{\boldsymbol{\beta}}_{ols}}$ 和 $\widehat{\mathbf{V}}^{\mathrm{HC3}}_{\hat{\boldsymbol{\beta}}_{ols}}$ 在数值上比较接近。$\widehat{\mathbf{V}}^{\mathrm{HC1}}_{\hat{\boldsymbol{\beta}}_{ols}}$ 是 Stata 默认的异方差稳健估计量；根据式 (3.10)，$\widehat{\mathbf{V}}^{\mathrm{HC2}}_{\hat{\boldsymbol{\beta}}_{ols}}$ 在球形扰动假设下具有无偏性；根据式 (3.11)，$\widehat{\mathbf{V}}^{\mathrm{HC3}}_{\hat{\boldsymbol{\beta}}_{ols}}$ 提供了三个估计量中较谨慎的异方差稳健估计量。此外，注意到 $\tilde{e}_i = (1-h_{ii})^{-1}e_i$ 和 $\bar{e}_i = (1-h_{ii})^{-1/2}e_i$，[3] 当至少有一个观测具有较大的杠杆值时，三个估计量给出的结果可能存在较大的差异。

3.4 异方差自相关稳健标准误

自相关序列样本均值的渐近理论

若时间序列存在自相关，其样本均值的渐近方差表达式中将包含自协方差项。考虑一元时间序列 $\{y_t\}_{t=1}^T$，定义 $\bar{y} = (1/T)\sum_{t=1}^T y_t$，则有：

$$\begin{aligned}
\mathrm{Var}(\sqrt{T}\bar{y}) &= \frac{1}{T}\mathrm{Var}\left(\sum_{t=1}^{T} y_t\right) = \frac{1}{T}\left(T\gamma_0 + \sum_{t=1}^{T}\sum_{s\neq t}^{T}\mathrm{Cov}(y_t, y_s)\right) \\
&= \gamma_0 + \frac{2}{T}\sum_{t=1}^{T}\sum_{s<t}^{T}\mathrm{Cov}(y_t, y_s) = \gamma_0 + \frac{2}{T}\sum_{j=1}^{T-1}(T-j)\gamma_j \\
&= \gamma_0 + 2\sum_{j=1}^{T-1}(1-\frac{j}{T})\gamma_j.
\end{aligned}$$

[2] 回归中一般都有常数项，此时，$h_{ii} \geqslant 1/n$。

[3] 见本章附录 2。

定理 B.8 给出了平稳遍历鞅差序列的中心极限定理，若时间序列虽满足平稳遍历性但存在自相关，我们需要施加额外的假设保证中心极限定理依然成立。比较常用的假设是遍历平稳过程的 Gordin 条件（见 White (2001) 的第五章）：

(a) $\mathbb{E}(y_t^2) < \infty$；

(b) 当 $j \to \infty$ 时，$\mathbb{E}(y_t \mid y_{t-j}, y_{t-j-1}, \ldots) \underset{m.s.}{\to} 0$；

(c) $\sum_{j=0}^{\infty}[\mathbb{E}(r_{tj}^2)]^{1/2} < \infty$，其中，$r_{tj} \equiv \mathbb{E}(y_t \mid \mathscr{F}_{t-j}) - \mathbb{E}(y_t \mid \mathscr{F}_{t-j-1})$，$\mathscr{F}_t \equiv \{y_t, y_{t-1}, \ldots\}$。

注 条件 (a) 要求 y_t 具有有限的二阶矩，由于严平稳过程的二阶矩可能不存在，所以，(a) 是对平稳性施加的额外限制；条件 (b) 可以推出 $\mathbb{E}(y_t) = 0$，直观上，随着 j 趋于无穷，对 y_t 的预测使用的信息越来越少，直至最后接近于不使用任何信息的预测，即，无条件期望；为了更好地理解条件 (c)，将 y_t 表示为：

$$\begin{aligned} y_t &= [y_t - \mathbb{E}(y_t \mid \mathscr{F}_{t-1})] + [\mathbb{E}(y_t \mid \mathscr{F}_{t-1}) - \mathbb{E}(y_t \mid \mathscr{F}_{t-2})] + \cdots \\ &\quad + [\mathbb{E}(y_t \mid \mathscr{F}_{t-j+1}) - \mathbb{E}(y_t \mid \mathscr{F}_{t-j})] + \mathbb{E}(y_t \mid \mathscr{F}_{t-j}) \\ &= (r_{t0} + r_{t1} + \cdots + r_{t,j-1}) + \mathbb{E}(y_t \mid \mathscr{F}_{t-j}). \end{aligned}$$

由条件 (b) 可知，对任意的 t，当 $j \to \infty$ 时，$y_t - (r_{t0} + r_{t1} + \cdots + r_{t,j-1})$ 依均方收敛到 0，因此，

$$y_t = \sum_{j=1}^{\infty} r_{tj}.$$

注意到 r_{tj} 衡量的是信息由 $\mathscr{F}_{t-j-1}$ 增加到 $\mathscr{F}_{t-j}$ 后对 y_t 预测的改进，y_t 可以表示为当前时刻和所有过去时刻的改进之和。条件 (c) 直观上要求距离当前时刻很久之前的改进对 y_t 没有显著的影响，这实际上是对 $\{y_t\}$ 自相关强度的限制。

满足 Gordin 条件的平稳遍历过程有如下结论成立：

定理 3.5. 零均值平稳遍历过程的中心极限定理

假设 $\{y_t\}$ 是平稳遍历过程且满足 Gordin 条件，则有：

- $\mathbb{E}(y_t) = 0$；
- 自协方差 $\{\gamma_j\}$ 绝对可加；
- $\sqrt{T}\,\bar{y} \underset{d}{\to} \mathcal{N}\left(0, \sum_{j=-\infty}^{\infty} \gamma_j\right)$.

♡

Gordin 条件可以容易地推广到多元情形：

(a) $\mathbb{E}(\mathbf{y}_t\mathbf{y}_t')$ 存在且有限；

(b) 当 $j \to \infty$ 时，$\mathbb{E}(\mathbf{y}_t \mid \mathbf{y}_{t-j}, \mathbf{y}_{t-j-1}, \ldots) \underset{m.s.}{\to} \mathbf{0}$；

(c) $\sum_{j=0}^{\infty}[\mathbb{E}(\mathbf{r}_{tj}'\mathbf{r}_{tj})]^{1/2} < \infty$，其中，

$$\mathbf{r}_{tj} \equiv \mathbb{E}(\mathbf{y}_t \mid \mathbf{y}_{t-j}, \mathbf{y}_{t-j-1}, \ldots) - \mathbb{E}(\mathbf{y}_t \mid \mathbf{y}_{t-j-1}, \mathbf{y}_{t-j-2}, \ldots).$$

满足多元 Gordin 条件的向量平稳遍历过程有如下结论成立：

定理 3.6. 零均值向量平稳遍历过程的中心极限定理

假设 $\{\mathbf{y}_t\}$ 是向量平稳遍历过程且满足多元 Gordin 条件，则有：

- $\mathbb{E}(\mathbf{y}_t) = 0$；
- 自协方差矩阵 $\{\mathbf{\Gamma}_j\}$ 绝对可加，其中，$\mathbf{\Gamma}_j \equiv \mathbb{E}[\mathbf{y}_t\mathbf{y}_{t-j}']$；
- $\sqrt{T}\,\bar{\mathbf{y}} \xrightarrow[d]{} \mathcal{N}\left(\mathbf{0}, \sum_{j=-\infty}^{\infty}\mathbf{\Gamma}_j\right)$.

♡

$\bar{\mathbf{y}}$ 的渐近方差 $\sum_{j=-\infty}^{\infty}\mathbf{\Gamma}_j$ 称为长期方差（long-run variance）。

长期方差的估计

假设 2.5* 要求 $\{\mathbf{g}_t\}$ 是鞅差序列，由于鞅差序列不存在自相关，$\bar{\mathbf{g}}$ 的渐近方差为 $\mathbf{S} = \mathbb{E}(\mathbf{g}_t\mathbf{g}_t')$。事实上，

$$\mathrm{Avar}(\bar{\mathbf{g}}) = \mathrm{Var}(\sqrt{n}\bar{\mathbf{g}}) = \mathrm{Var}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathbf{g}_i\right) = \frac{1}{n}\sum_{i=1}^{n}\mathrm{Var}(\mathbf{g}_i) = \mathbf{S}.$$

我们可以通过如下假设放松无自相关的限制：

假设 2.5′：（Gordin 条件）$\{\mathbf{g}_t\}$ 满足多元 Gordin 条件，其长期方差是非奇异的。

由定理 3.6，$\sqrt{n}\bar{\mathbf{g}}$ 收敛到正态分布，其渐近方差 $\mathbf{S}$ 为：

$$\mathbf{S} = \sum_{j=-\infty}^{\infty}\mathbf{\Gamma}_j = \mathbf{\Gamma}_0 + \sum_{j=1}^{\infty}(\mathbf{\Gamma}_j + \mathbf{\Gamma}_j'), \quad \text{其中，} \mathbf{\Gamma}_j = \mathbb{E}(\mathbf{g}_t\mathbf{g}_{t-j}'),$$

这里我们用到了 $\mathbf{\Gamma}_{-j} = \mathbf{\Gamma}_j'$。只要我们可以一致地估计长期方差 $\mathbf{S}$，定理 2.9 和定理 2.11 的结论依然成立。同时，因为 $\{\mathbf{\Gamma}_j\}$ 绝对可加，所以，随着 $j \to \infty$，$\mathbf{\Gamma}_j \to \mathbf{0}$。我们可以选取足够大的正整数 q，使得

$$\mathbf{S}_q = \sum_{j=-q}^{q}\mathbf{\Gamma}_j$$

可以作为 $\mathbf{S}$ 的近似。这里的 q 在文献中被称为滞后截断数（lag truncation number）或窗宽（bandwidth）。

对于固定的 j，$\mathbf{\Gamma}_j$ 可以用如下样本自协方差估计：

$$\widehat{\mathbf{\Gamma}}_j = \frac{1}{n}\sum_{t=j+1}^{n}\hat{\mathbf{g}}_t\hat{\mathbf{g}}_{t-j}', \quad j = 0, 1, \ldots, n-1,$$

其中，$\hat{\mathbf{g}}_t = \mathbf{x}_t e_t$。在适当的矩条件下，可以证明 $\widehat{\mathbf{\Gamma}}_j$ 是 $\mathbf{\Gamma}_j$ 的一致估计量。因此，对任意给定的 $q > 0$，

$$\hat{\mathbf{S}}_q = \sum_{j=-q}^{q}\widehat{\mathbf{\Gamma}}_j$$

是 $\mathbf{S}_q$ 的一致估计。

如果 $\{\mathbf{g}_t\}$ 满足：对已知且有限的正整数 q，当 $j > q$ 时，$\mathbf{\Gamma}_j = \mathbf{0}$，则有 $\mathbf{S} = \mathbf{S}_q$。于是，$\hat{\mathbf{S}}_q$ 是 $\mathbf{S}$ 的一致估计。更一般地，q 可以随着 n 趋于无穷，只要 q 趋于无穷的速度足够慢，$\hat{\mathbf{S}}_q$ 仍然是 $\mathbf{S}$ 的一致估计（见 White et al. (1984)）。若 q 是未知的，实际中需要选择适当的 q 对长期方差进行截断，这会导致两个问题：一是估计的结果对 q 的选择敏感；二是 $\hat{\mathbf{S}}_q$ 可能不是半正定的。例如，假设 $\mathbf{g}_t$ 是标量，若 $q = 1$，$\hat{\mathbf{S}}_1 = \hat{\gamma}_0(1 + 2\hat{\rho}_1)$，当 $\hat{\rho}_1 < 0.5$ 时，

$\hat{\mathbf{S}}_1$ 为负数。为了解决这两个问题，Newey et al. (1987) 提出了如下 Newey-West 估计量：

$$\hat{\mathbf{S}}_{nw} = \sum_{j=-q}^{q} \left(1 - \frac{|j|}{q+1}\right) \widehat{\mathbf{\Gamma}}_j. \tag{3.12}$$

上式实际上是 $\{\widehat{\mathbf{\Gamma}}_j\}$ 的加权求和，其权重函数称为 Bartlett 核函数。基于 $\hat{\mathbf{S}}_{nw}$ 的最小二乘估计量的协方差矩阵估计称为异方差自相关一致（heteroskedasticity and autocorrelation consistent, HAC）协方差估计或异方差自相关稳健（heteroskedasticity and autocorrelation robust, HAR）协方差估计。为了使 $\hat{\mathbf{S}}_{nw}$ 是 $\mathbf{S}$ 的一致估计量，q 需要随着 n 趋于无穷，Hansen (1992) 指出 q 趋于无穷的速度要不快于 $n^{1/3}$。Andrews (1991) 指出最小化 $\hat{\mathbf{S}}_{nw}$ 的均方误差的最优窗宽 q 满足 $q = Cn^{1/3}$，其中，C 是与自协方差有关的常数，选择最优窗宽的 Andrews 规则为：

$$q = \left[6\frac{\rho^2}{(1-\rho^2)^2}\right]^{1/3} n^{1/3},$$

其中，ρ 是自相关参数，若 $\mathbf{g}_t$ 是标量，ρ 是 $\mathbf{g}_t$ 的一阶自相关系数。实际使用时，ρ 也可以取成某些默认值，例如，$\rho = 0.5$，此时，$q = 1.4n^{1/3}$。

注 在 Stata 中，可以对时间序列数据使用命令 newey 进行 OLS 回归并汇报系数估计的 Newey-West 标准误，在使用此命令前要先使用命令 tsset 声明时间变量。newey 命令的使用格式为：newey *depvar indepvars*, lag(q)，其中，*depvar* 为被解释变量，*indepvars* 为解释变量，lag(q) 设置最大的滞后阶数为 q。

3.5 聚类标准误

前文讨论的各个观测之间的相关主要是指时间序列数据的自相关。对于横截面数据，我们通常假设各个观测之间是彼此独立的。但是，横截面个体之间也可能存在相关性，例如，国内各个地级市之间有经济上的往来，利用地级市数据研究经济问题时假设横截面独立就不符合实际情况；股票价格会受到经济金融政策、突发事件等共同因素的冲击，也会导致横截面的相关性。与时间序列的自相关不同，横截面的相关性没有一个像时间一样的自然顺序，这导致为横截面相关性建模往往比较困难。空间计量（见 Anselin (1988)）通过定义“相关矩阵”在计量模型中显示地刻画横截面相关，这种相关矩阵的结构通常由横截面个体的地理距离或经济意义上的“距离”决定。在面板数据中，可以通过引入共同因子结构对横截面相关性进行建模（见 Pesaran (2006) 和 Bai (2009)）。关于使用空间计量和因子结构进行横截面相关性建模的较详细介绍可以参考 Xu et al. (2016)。本小节并不讨论为横截面相关性建模的问题，而是讨论在存在横截面相关性的情况下如何做稳健的统计推断。这里考虑的样本为聚类样本（clustered sample）。在聚类样本中，各个观测个体被划分为有限个聚类，各个聚类之间彼此独立或不相关，但是允许同一聚类内部的个体存在任意形式的相关性。[4]

具体地，我们考虑样本 $(y_{ig}, \mathbf{x}_{ig})$，其中，$g = 1, \ldots, G$ 是表示 G 个不同聚类的下标；$i = 1, \ldots, n_g$ 是第 g 个聚类内观测个体的下标。根据这两个下标的定义，样本总量为：

[4] Cameron et al. (2015) 和 MacKinnon et al. (2022a) 对聚类标准误进行了详细地讨论，推荐想进一步了解相关内容的读者阅读。

$n=\sum_{g=1}^{G} n_g$。令 $\mathbf{y}_g=(y_{1g},\ldots,y_{n_g g})'$ 表示 $n_g\times 1$ 维第 g 个聚类中的所有被解释变量观测值构成的向量；$\mathbf{X}_g=(\mathbf{x}_{1g},\ldots,\mathbf{x}_{n_g g})'$ 表示 $n_g\times K$ 维第 g 个聚类中的所有解释变量观测值构成的数据矩阵。个体层面的线性回归方程可以表示为：

$$y_{ig}=\mathbf{x}_{ig}'\boldsymbol{\beta}+\varepsilon_{ig},\quad i=1,\ldots,n_g;\ g=1,\ldots,G.$$

利用刚刚引入的记号，聚类层面的线性回归方程可以表示为：

$$\mathbf{y}_g=\mathbf{X}_g\boldsymbol{\beta}+\boldsymbol{\varepsilon}_g,\quad g=1,\ldots,G,$$

其中，$\boldsymbol{\varepsilon}_g=(\varepsilon_{1g},\ldots,\varepsilon_{n_g g})'$ 是第 g 个聚类中的 $n_g\times 1$ 维的误差向量。显然，把所有的 $\mathbf{X}_g$ 按列堆叠在一起便得到 $n\times K$ 维的数据矩阵 $\mathbf{X}$。类似地，我们也可以通过 $\mathbf{y}_g$ 和 $\boldsymbol{\varepsilon}_g$ 的堆叠分别得到 $\mathbf{y}$ 和 $\boldsymbol{\varepsilon}$。因此，假设 2.1 中的 $\mathbf{y}=\mathbf{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon}$ 依然成立。由式 (2.8)，

$$\begin{aligned}\hat{\boldsymbol{\beta}}_{ols} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}=\left(\sum_{g=1}^{G}\sum_{i=1}^{n_g}\mathbf{x}_{ig}\mathbf{x}_{ig}'\right)^{-1}\left(\sum_{g=1}^{G}\sum_{i=1}^{n_g}\mathbf{x}_{ig}y_{ig}\right)\\ &= \left(\sum_{g=1}^{G}\mathbf{X}_g'\mathbf{X}_g\right)^{-1}\left(\sum_{g=1}^{G}\mathbf{X}_g'\mathbf{y}_g\right).\end{aligned}\tag{3.13}$$

个体层面的残差可以表示为 $e_{ig}=y_{ig}-\mathbf{x}_{ig}'\hat{\boldsymbol{\beta}}_{ols}$；聚类层面的残差向量可以表示为 $\boldsymbol{e}_g=\mathbf{y}_g-\mathbf{X}_g\hat{\boldsymbol{\beta}}_{ols}$。注意到，由式 (2.17)，

$$\hat{\boldsymbol{\beta}}_{ols}-\boldsymbol{\beta}=\left(\sum_{g=1}^{G}\mathbf{X}_g'\mathbf{X}_g\right)^{-1}\left(\sum_{g=1}^{G}\mathbf{X}_g'\boldsymbol{\varepsilon}_g\right)\equiv\left(\sum_{g=1}^{G}\mathbf{X}_g'\mathbf{X}_g\right)^{-1}\left(\sum_{g=1}^{G}\boldsymbol{s}_g\right),\tag{3.14}$$

其中，$\boldsymbol{s}_g=\mathbf{X}_g'\boldsymbol{\varepsilon}_g$，MacKinnon et al. (2022a) 将 $\boldsymbol{s}_g$ 称为第 g 个聚类的得分（score）向量。对于得分向量，我们需要施加如下假设：

假设 3.1：（聚类规则）

(a) $\mathbb{E}(\boldsymbol{s}_g)=\mathbf{0}$，$g=1,\ldots,G$；

(b) $\mathbb{E}(\boldsymbol{s}_g\boldsymbol{s}_g')=\boldsymbol{\Sigma}_g$，$\mathbb{E}(\boldsymbol{s}_g\boldsymbol{s}_{g^*}')=\mathbf{0}$，$g,\ g^*=1,\ldots,G$，$g\neq g^*$。

注 在 (b) 中，我们要求不同聚类的得分向量彼此之间不相关；但是，$\boldsymbol{\Sigma}_g$ 可以具有任意形式的异方差和（或）聚类内的相关结构。使假设 3.1 成立的一组充分条件为：

假设 3.1*：（聚类的外生性和条件不相关）

(a) $\mathbb{E}(\boldsymbol{\varepsilon}_g\mid\mathbf{X})=\mathbf{0}$，$g=1,\ldots,G$；

(b) $\mathbb{E}(\boldsymbol{\varepsilon}_g\boldsymbol{\varepsilon}_g'\mid\mathbf{X})=\boldsymbol{\Omega}_g$，$\mathbb{E}(\boldsymbol{\varepsilon}_g\boldsymbol{\varepsilon}_{g^*}'\mid\mathbf{X})=\mathbf{0}$，$g,\ g^*=1,\ldots,G$，$g\neq g^*$。

事实上，根据假设 3.1* 中的条件 (a)，

$$\mathbb{E}(\boldsymbol{s}_g)=\mathbb{E}[\mathbb{E}(\mathbf{X}_g'\boldsymbol{\varepsilon}_g\mid\mathbf{X})]=\mathbb{E}[\mathbf{X}_g'\mathbb{E}(\boldsymbol{\varepsilon}_g\mid\mathbf{X})]=\mathbf{0}.$$

此外，根据假设 3.1* 中的条件 (b)，可以推出：

$$\boldsymbol{\Sigma}_g=\mathbb{E}(\boldsymbol{s}_g\boldsymbol{s}_g')=\mathbb{E}[\mathbb{E}(\boldsymbol{s}_g\boldsymbol{s}_g'\mid\mathbf{X})]=\mathbb{E}[\mathbf{X}_g'\mathbb{E}(\boldsymbol{\varepsilon}_g\boldsymbol{\varepsilon}_g'\mid\mathbf{X})\mathbf{X}_g]=\mathbb{E}[\mathbf{X}_g'\boldsymbol{\Omega}_g\mathbf{X}_g],$$

和

$$\mathbb{E}(\boldsymbol{s}_g\boldsymbol{s}_{g^*}')=\mathbb{E}[\mathbb{E}(\boldsymbol{s}_g\boldsymbol{s}_{g^*}'\mid\mathbf{X})]=\mathbb{E}[\mathbf{X}_g'\mathbb{E}(\boldsymbol{\varepsilon}_g\boldsymbol{\varepsilon}_{g^*}'\mid\mathbf{X})\mathbf{X}_{g^*}]=\mathbf{0}.$$

若假设 $\mathbb{E}(\varepsilon_{ig}^2\mid\mathbf{X})=\sigma^2$ 和 $\mathbb{E}(\varepsilon_{ig}\varepsilon_{jg}\mid\mathbf{X})=0$，$i\neq j$，则有：$\boldsymbol{\Omega}_g=\sigma^2\mathbf{I}_{n_g}$，并

且 $\mathbb{E}(\boldsymbol{s}_g\boldsymbol{s}_g' \mid \mathbf{X}) = \sigma^2\mathbf{X}_g'\mathbf{X}_g$。此时，容易验证最小二乘估计量的条件方差依然具有定理 2.2 中结论 (b) 的形式，即，$\mathrm{Var}(\hat{\boldsymbol{\beta}}_{ols} \mid \mathbf{X}) = \sigma^2\left(\mathbf{X}'\mathbf{X}\right)^{-1}$。然而，一般情况下，由假设 3.1* 和式 (3.14)，我们有：

$$\mathrm{Var}(\hat{\boldsymbol{\beta}}_{ols} \mid \mathbf{X}) = \mathrm{Var}(\hat{\boldsymbol{\beta}}_{ols} - \boldsymbol{\beta} \mid \mathbf{X}) = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\left(\sum_{g=1}^{G}\mathbf{X}_g'\boldsymbol{\Omega}_g\mathbf{X}_g\right)\left(\mathbf{X}'\mathbf{X}\right)^{-1}. \tag{3.15}$$

式 (3.15) 与 $\sigma^2\left(\mathbf{X}'\mathbf{X}\right)^{-1}$ 存在区别的原因是 $\mathbb{E}(\boldsymbol{\varepsilon}_g\boldsymbol{\varepsilon}_g' \mid \mathbf{X})$ 的条件异方差和聚类内的横截面相关。为了看清楚聚类内的横截面相关对最小二乘估计量的条件方差的影响，考虑如下情况：

$$\mathbb{E}(\varepsilon_{ig}^2 \mid \mathbf{X}) = \sigma^2; \quad \mathbb{E}\left[\varepsilon_{ig}\varepsilon_{jg}|\mathbf{X}\right] = \sigma^2\rho, \quad i \neq j; \quad n_g = N, \quad g = 1, \ldots, G.$$

此外，假设同一聚类内的解释变量 $\mathbf{x}_{ig}$ 的取值都相同。容易验证，此时的最小二乘估计量的条件方差为：

$$\mathrm{Var}(\hat{\boldsymbol{\beta}}_{ols} \mid \mathbf{X}) = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\sigma^2(1+\rho(N-1)).$$

从上式可以看出，在聚类内存在横截面相关的情况下，最小二乘估计量的条件方差与球形扰动假设下的古典形式的区别依赖于聚类内的样本数量和聚类内横截面个体间的相关程度。因此，即使聚类内横截面个体间的相关性很弱，大的聚类内样本量也会导致 $\hat{\boldsymbol{\beta}}_{ols}$ 的真实条件方差与 $\sigma^2\left(\mathbf{X}'\mathbf{X}\right)^{-1}$ 之间存在较大差异。例如，当 $\rho = 0.1$ 和 $N = 50$ 时，真实的条件方差是 $\sigma^2\left(\mathbf{X}'\mathbf{X}\right)^{-1}$ 的 5.9 倍。

聚类稳健方差估计量

为了估计式 (3.15) 中的 $\mathrm{Var}(\hat{\boldsymbol{\beta}}_{ols} \mid \mathbf{X})$，我们可以将 $\boldsymbol{\Omega}_g$ 替换为 $\boldsymbol{e}_g\boldsymbol{e}_g'$，其中，$\boldsymbol{e}_g = \left(e_{1g}, \ldots, e_{n_g g}\right)'$，表示第 g 个聚类的回归残差构成的列向量，从而得到如下估计量：

$$\left(\mathbf{X}'\mathbf{X}\right)^{-1}\left(\sum_{g=1}^{G}\mathbf{X}_g'\boldsymbol{e}_g\boldsymbol{e}_g'\mathbf{X}_g\right)\left(\mathbf{X}'\mathbf{X}\right)^{-1}.$$

在实际使用时，一般会对上式做小样本调整，Stata 中 reg 命令计算聚类稳健标准误时使用的聚类稳健协方差矩阵估计量为：

$$\widehat{\mathbf{V}}_{\hat{\boldsymbol{\beta}}_{ols}}^{\mathrm{CV1}} = \left(\frac{n-1}{n-K}\right)\left(\frac{G}{G-1}\right)\left(\mathbf{X}'\mathbf{X}\right)^{-1}\left(\sum_{g=1}^{G}\mathbf{X}_g'\boldsymbol{e}_g\boldsymbol{e}_g'\mathbf{X}_g\right)\left(\mathbf{X}'\mathbf{X}\right)^{-1}. \tag{3.16}$$

当 $G = n$ 时，$\widehat{\mathbf{V}}_{\hat{\boldsymbol{\beta}}_{ols}}^{\mathrm{CV1}}$ 退化为式 (3.6) 中的 $\widehat{\mathbf{V}}_{\hat{\boldsymbol{\beta}}_{ols}}^{\mathrm{HC1}}$。使用聚类稳健标准误进行统计推断的 Stata 命令格式为：reg *depvar indepvars*, vce(cluster *clustvar*)，其中，*clustvar* 为识别不同聚类的变量。

与本章附录 2 中关于残差的讨论类似，$\boldsymbol{e}_g\boldsymbol{e}_g'$ 并不是 $\boldsymbol{\Omega}_g$ 的最好估计，Bell et al. (2002) 提出了另外两种聚类稳健协方差估计量：

$$\widehat{\mathbf{V}}_{\hat{\boldsymbol{\beta}}_{ols}}^{\mathrm{CV2}} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\left(\sum_{g=1}^{G}\mathbf{X}_g'\bar{\boldsymbol{e}}_g\bar{\boldsymbol{e}}_g'\mathbf{X}_g\right)\left(\mathbf{X}'\mathbf{X}\right)^{-1}, \tag{3.17}$$

$$\widehat{\mathbf{V}}_{\hat{\boldsymbol{\beta}}_{ols}}^{\mathrm{CV3}} = \left(\frac{G-1}{G}\right)(\mathbf{X}'\mathbf{X})^{-1}\left(\sum_{g=1}^{G}\mathbf{X}_g'\tilde{\boldsymbol{e}}_g\tilde{\boldsymbol{e}}_g'\mathbf{X}_g\right)(\mathbf{X}'\mathbf{X})^{-1}, \tag{3.18}$$

其中，$\bar{\boldsymbol{e}}_g = \mathbf{M}_{gg}^{-1/2}\boldsymbol{e}_g$，$\tilde{\boldsymbol{e}}_g = \mathbf{M}_{gg}^{-1}\boldsymbol{e}_g$，$\mathbf{M}_{gg} = \mathbf{I}_{n_g} - \mathbf{X}_g(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_g'$。当 $G = n$ 时，$\widehat{\mathbf{V}}_{\hat{\boldsymbol{\beta}}_{ols}}^{\mathrm{CV2}}$ 退化为式 (3.7) 中的 $\widehat{\mathbf{V}}_{\hat{\boldsymbol{\beta}}_{ols}}^{\mathrm{HC2}}$。

Stata 的非官方命令 reg_sandwich 提供了基于 (3.17) 计算的聚类稳健标准误。使用此命令前，需要先进行安装，命令如下：

```
. ssc install reg_sandwich, replace
```

reg_sandwich 命令的使用格式为：reg_sandwich *depvar indepvars*, cluster(*clustvar*)。关于此命令背后计量理论的详细介绍读者可以参考 Pustejovsky et al. (2018)。

当 n_g 比较小时，根据式 (3.18) 计算 $\widehat{\mathbf{V}}_{\hat{\boldsymbol{\beta}}_{ols}}^{\mathrm{CV3}}$ 是可行的。然而，当存在着某一个聚类 g 具有很大的 n_g 时，基于 (3.18) 的计算会出现问题，这是因为计算过程中要对每一个聚类的 $n_g \times n_g$ 维矩阵 $\mathbf{M}_{gg}$ 进行求逆并保存结果。为了解决这一问题，MacKinnon et al. (2022c) 提供了一种基于 Jackknife 方法的高效算法[5]，并开发了 Stata 的非官方命令 summclust 实现此算法。使用此命令前，也需要先进行安装，命令如下：

```
. ssc install summclust, replace
```

summclust 命令的使用格式为：summclust *core_indepvar*, yvar(*depvar*) xvar(*other_indepvars*) cluster(*clustvar*)，其中，*core_indepvar* 是核心解释变量，*other_indepvars* 为其他控制变量。summclust 命令的输出结果主要包括了核心解释变量的系数估计以及基于式 (3.16) 和 (3.18) 的聚类稳健标准误。

统计模拟显示，与 Stata 回归命令默认使用的 $\widehat{\mathbf{V}}_{\hat{\boldsymbol{\beta}}_{ols}}^{\mathrm{CV1}}$ 相比，在小样本情况下，使用由 $\widehat{\mathbf{V}}_{\hat{\boldsymbol{\beta}}_{ols}}^{\mathrm{CV3}}$ 计算的标准误进行统计推断得到的结论更加可靠。因此，MacKinnon et al. (2022c) 推荐使用 $\widehat{\mathbf{V}}_{\hat{\boldsymbol{\beta}}_{ols}}^{\mathrm{CV3}}$ 进行统计推断。

3.6 重抽样方法

古典模型的统计推断使用了检验统计量的精确分布，但是，保证精确分布成立需要同方差和正态性等过于严格的假设；另一方面，大样本统计推断基于统计量的渐近分布，因此，在样本量十分有限的情况下，统计量的真实分布与渐近分布之间可能存在较大的差异。本节讨论在实证分析中被广泛使用的重抽样方法，其中包括 Jackknife 和 Bootstrap。当样本量为 n 时，Jackknife 通过 n 个“去一”估计量的经验分布分析全样本估计量的统计性质，通常用于方差的估计；Bootstrap 通过对原数据集进行有放回的随机抽样得到样本量同样为 n 的 Bootstrap 样本并基于此样本得到参数的估计，如此重复多次最终得到参数估计量的抽样分布的估计。Bootstrap 常用于方差估计、构造置信区间和假设检验。Hansen (2022a) 的第十章以及 Efron et al. (2016) 的第十章和第十一章对这两种重抽样方

[5] 我们将在下一节讨论 Jackknife 方法。

法的实施步骤和理论性质进行了详细地介绍。

Jackknife

Jackknife 利用“去一”估计量的经验分布得到全样本估计量的各阶矩的估计。具体地，令 $\hat{\boldsymbol{\theta}}$ 是向量值参数 $\boldsymbol{\theta}$ 的任意估计量，即，$\hat{\boldsymbol{\theta}}$ 是样本量为 n 的简单随机样本的一个函数；再令 $\mathbf{V}_{\hat{\boldsymbol{\theta}}} = \mathrm{Var}(\hat{\boldsymbol{\theta}})$ 表示 $\hat{\boldsymbol{\theta}}$ 的方差。$\boldsymbol{\theta}$ 的“去一”估计量 $\hat{\boldsymbol{\theta}}_{(-i)}$ 利用了 $\hat{\boldsymbol{\theta}}$ 的函数形式，但是，在估计 $\boldsymbol{\theta}$ 时，$\hat{\boldsymbol{\theta}}_{(-i)}$ 未使用第 i 个样本观测。$\mathbf{V}_{\hat{\boldsymbol{\theta}}}$ 的 Jackknife 估计量定义为：

$$\widehat{\mathbf{V}}_{\hat{\boldsymbol{\theta}}}^{\text{jack}} = \frac{n-1}{n}\sum_{i=1}^{n}\left(\hat{\boldsymbol{\theta}}_{(-i)} - \bar{\boldsymbol{\theta}}\right)\left(\hat{\boldsymbol{\theta}}_{(-i)} - \bar{\boldsymbol{\theta}}\right)', \tag{3.19}$$

其中，$\bar{\boldsymbol{\theta}}$ 是“去一”估计量的样本均值，即，

$$\bar{\boldsymbol{\theta}} = \frac{1}{n}\sum_{i=1}^{n}\hat{\boldsymbol{\theta}}_{(-i)}.$$

$\hat{\boldsymbol{\theta}}$ 中第 k 个元素的 Jackknife 标准误是 $\widehat{\mathbf{V}}_{\hat{\boldsymbol{\theta}}}^{\text{jack}}$ 的对角线上第 k 个元素的平方根。

从形式上看，式 (3.19) 类似于样本协方差矩阵的公式，但是，求和符号前的 $(n-1)/n$ 并不直观。为了更好地理解式 (3.19)，考虑如下例子：令 $\hat{\theta}$ 是 $(x_1, \ldots, x_n)$ 的样本均值，我们有：

$$\begin{aligned}
\hat{\theta}_{(-i)} &= \frac{1}{n-1}\sum_{j\neq i}x_j = \frac{n}{n-1}\bar{x} - \frac{1}{n-1}x_i, \quad \bar{x} = \frac{1}{n}\sum_{i=1}^{n}x_i, \\
\bar{\theta} &= \frac{1}{n}\sum_{i=1}^{n}\hat{\theta}_{(-i)} = \bar{x}, \quad \hat{\theta}_{(-i)} - \bar{\theta} = \frac{1}{n-1}(\bar{x} - x_i).
\end{aligned}$$

于是，

$$\widehat{\mathbf{V}}_{\hat{\theta}}^{\text{jack}} = \frac{n-1}{n}\sum_{i=1}^{n}\left(\frac{1}{n-1}\right)^2(\bar{x} - x_i)^2 = \frac{1}{n}\left(\frac{1}{n-1}\right)\sum_{i=1}^{n}(\bar{x} - x_i)^2,$$

从上式可以看出，式 (3.19) 给出的 Jackknife 估计量即为 $\bar{x}$ 的方差的估计量。

接下来，我们考虑最小二乘估计量 $\hat{\boldsymbol{\beta}}_{ols}$ 的方差的 Jackknife 估计。由式 (3.41)，“去一”估计量的样本均值为：

$$\bar{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}_{ols} - \left(\mathbf{X}'\mathbf{X}\right)^{-1}\tilde{\boldsymbol{\mu}}, \quad \tilde{\boldsymbol{\mu}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\tilde{e}_i.$$

将式 (3.41) 和 $\bar{\boldsymbol{\beta}}$ 代入式 (3.19) 得到：

$$\begin{aligned}
\widehat{\mathbf{V}}_{\hat{\boldsymbol{\beta}}_{ols}}^{\text{jack}} &= \frac{n-1}{n}\sum_{i=1}^{n}\left(\hat{\boldsymbol{\beta}}_{(-i)} - \bar{\boldsymbol{\beta}}\right)\left(\hat{\boldsymbol{\beta}}_{(-i)} - \bar{\boldsymbol{\beta}}\right)' \\
&= \frac{n-1}{n}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\left[\sum_{i=1}^{n}(\mathbf{x}_i\tilde{e}_i - \tilde{\boldsymbol{\mu}})(\mathbf{x}_i\tilde{e}_i - \tilde{\boldsymbol{\mu}})'\right]\left(\mathbf{X}'\mathbf{X}\right)^{-1} \\
&= \frac{n-1}{n}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\left(\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i'\tilde{e}_i^2 - n\tilde{\boldsymbol{\mu}}\tilde{\boldsymbol{\mu}}'\right)\left(\mathbf{X}'\mathbf{X}\right)^{-1} \\
&= \frac{n-1}{n}\widehat{\mathbf{V}}_{\hat{\beta}_{ols}}^{\text{HC3}} - (n-1)\left(\mathbf{X}'\mathbf{X}\right)^{-1}\tilde{\boldsymbol{\mu}}\tilde{\boldsymbol{\mu}}'\left(\mathbf{X}'\mathbf{X}\right)^{-1}.
\end{aligned} \tag{3.20}$$

由于向量 $\tilde{\boldsymbol{\mu}}$ 的各个元素的绝对值通常很小，因此，$\widehat{\mathbf{V}}_{\hat{\boldsymbol{\beta}}_{ols}}^{\text{jack}} \approx \widehat{\mathbf{V}}_{\hat{\boldsymbol{\beta}}_{ols}}^{\text{HC3}}$，即，回归系数的方差

的 Jackknife 估计量与异方差稳健方差估计量是近似相等的。使用 Jackknife 估计量的优点是我们无需知道渐近方差的具体形式。

对于最小二乘估计量的非线性函数的方差的 Jackknife 估计也有类似结论。具体地，令 $\boldsymbol{\theta} = \mathbf{a}(\boldsymbol{\beta})$，$\hat{\boldsymbol{\theta}} = \mathbf{a}(\hat{\boldsymbol{\beta}}_{ols})$，$\boldsymbol{\theta}$ 的“去一”估计量为：

$$
\begin{aligned}
\hat{\boldsymbol{\theta}}_{(-i)} &= \mathbf{a}(\hat{\boldsymbol{\beta}}_{(-i)}) \\
&= \mathbf{a}\left(\hat{\boldsymbol{\beta}}_{ols} - \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{x}_i\tilde{e}_i\right) && \text{(由式 (3.41))} \\
&\approx \hat{\boldsymbol{\theta}} - \mathbf{A}(\hat{\boldsymbol{\beta}}_{ols})'\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{x}_i\tilde{e}_i, \quad \left(\mathbf{a}\left(\hat{\boldsymbol{\beta}}_{ols} - \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{x}_i\tilde{e}_i\right) \text{ 在 } \hat{\boldsymbol{\beta}}_{ols} \text{ 处做一阶泰勒展开}\right)
\end{aligned}
$$

其中，$\mathbf{A}(\hat{\boldsymbol{\beta}}_{ols}) = \partial\mathbf{a}(\boldsymbol{\beta})/\partial\boldsymbol{\beta}' \mid_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}_{ols}}$。于是，$\hat{\boldsymbol{\theta}}$ 的 Jackknife 方差估计量为：

$$
\begin{aligned}
\widehat{\mathbf{V}}^{\text{jack}}_{\hat{\boldsymbol{\theta}}} &= \frac{n-1}{n}\sum_{i=1}^{n}\left(\hat{\boldsymbol{\theta}}_{(-i)} - \bar{\boldsymbol{\theta}}\right)\left(\hat{\boldsymbol{\theta}}_{(-i)} - \bar{\boldsymbol{\theta}}\right)' \\
&= \frac{n-1}{n}\mathbf{A}(\hat{\boldsymbol{\beta}}_{ols})'\left(\mathbf{X}'\mathbf{X}\right)^{-1}\left(\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i'\tilde{e}_i^2 - n\tilde{\boldsymbol{\mu}}\tilde{\boldsymbol{\mu}}'\right)\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{A}(\hat{\boldsymbol{\beta}}_{ols}) \\
&= \mathbf{A}(\hat{\boldsymbol{\beta}}_{ols})' \cdot \widehat{\mathbf{V}}^{\text{jack}}_{\hat{\boldsymbol{\beta}}_{ols}} \cdot \mathbf{A}(\hat{\boldsymbol{\beta}}_{ols}) \\
&\approx \mathbf{A}(\hat{\boldsymbol{\beta}}_{ols})' \cdot \widehat{\mathbf{V}}^{\text{HC3}}_{\hat{\boldsymbol{\beta}}_{ols}} \cdot \mathbf{A}(\hat{\boldsymbol{\beta}}_{ols}).
\end{aligned}
$$

由上式可以看出，$\hat{\boldsymbol{\theta}}$ 的方差的 Jackknife 估计量近似等于使用 Delta 方法 B.5 得到的渐近方差。这再一次表明方差的 Jackknife 估计提供了一种关于渐近方差的近似。

综上，Jackknife 是计算方差和标准误的既简单又灵活的统计方法。在通常情况下，Jackknife 方差的估计结果与渐近理论提供的渐近方差非常接近。使用 Jackknife 的优点是它可以在渐近方差或标准误表达式未知或难于计算的情况下提供渐近方差的近似。

例 3.5 再次考虑 Wooldridge (2010) 中例 4.1 使用的数据集 mroz.dta，我们将工资对数（lwage）对受教育年限（educ）和工作经验（exper）做回归，并分别汇报系数估计量的异方差稳健标准误和 Jackknife 标准误，使用的 Stata 程序如下：

```
. use mroz.dta, replace
. keep if inlf == 1                //只保留有工作的女性样本
. reg lwage educ exper, vce(hc3)  //使用异方差稳健标准误HC3
. reg lwage educ exper, vce(jackknife, nodots) //使用Jackknife标准误
. jackknife, nodots: reg lwage educ exper      //与上一行命令作用相同
```

回归的系数估计结果、异方差稳健标准误 HC3 以及 Jackknife 标准误由表 3.1 给出，从中可以看到，异方差稳健标准误 HC3 与 Jackknife 标准误十分接近。

表 3.1: 例 3.5 回归结果

变量名	系数估计量	异方差稳健标准误	Jackknife 标准误
educ	0.109	0.0134415	0.0134258
exper	0.016	0.0041081	0.0041033
_cons	-0.400	0.1841642	0.1839489

Stata 中的 jackknife 命令的使用格式为：jackknife *exp_list* [, options] : *command*，其中，

常用的选项（option）有：cluster(*varlist*) 和 mse。若未设置聚类变量，对数据集中的每一个观测 i，jackknife 命令都利用去掉第 i 个观测之后的数据集执行一次 *command*，计算和保存 *exp_list* 设定的估计量或统计量，并基于保存的结果计算其 Jackknife 标准误；若设置了聚类变量，jackknife 命令将计算聚类的 Jackknife 标准误（详见下一小节的讨论）；若设置了 mse，Jackknife 方差估计将使用如下公式：

$$\widehat{\mathbf{V}}_{\hat{\boldsymbol{\theta}}}^{\text{jack,mse}} = \frac{n-1}{n}\sum_{i=1}^{n}\left(\hat{\boldsymbol{\theta}}_{(-i)}-\hat{\boldsymbol{\theta}}\right)\left(\hat{\boldsymbol{\theta}}_{(-i)}-\hat{\boldsymbol{\theta}}\right)'.$$

另外，默认的 *exp_list* 是 _b，即，最小二乘系数估计量，此时建议使用 vce(jackknife)。

聚类数据的 Jackknife

回忆式 (3.13) 给出的最小二乘估计量的聚类公式：

$$\hat{\boldsymbol{\beta}}_{ols} = \left(\sum_{g=1}^{G}\mathbf{X}_g'\mathbf{X}_g\right)^{-1}\left(\sum_{g=1}^{G}\mathbf{X}_g'\mathbf{y}_g\right).$$

与“去一”估计量类似，在计算聚类数据的 Jackknife 方差估计量时，我们考虑如下“删除聚类”估计量：

$$\hat{\boldsymbol{\beta}}_{(-g)} = \hat{\boldsymbol{\beta}}_{ols} - \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}_g'\tilde{\boldsymbol{e}}_g, \tag{3.21}$$

其中，

$$\tilde{\boldsymbol{e}}_g = \left(\mathbf{I}_{n_g} - \mathbf{X}_g\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}_g'\right)^{-1}\boldsymbol{e}_g.$$

于是，仿照式 (3.19)，最小二乘估计量 $\hat{\boldsymbol{\beta}}_{ols}$ 的方差的聚类稳健 Jackknife 估计量为：

$$\widehat{\mathbf{V}}_{\hat{\boldsymbol{\beta}}_{ols}}^{\text{jack,cluster}} = \frac{G-1}{G}\sum_{g=1}^{G}\left(\hat{\boldsymbol{\beta}}_{(-g)}-\bar{\boldsymbol{\beta}}\right)\left(\hat{\boldsymbol{\beta}}_{(-g)}-\bar{\boldsymbol{\beta}}\right)',$$

其中，$\bar{\boldsymbol{\beta}} = (1/G)\sum_{g=1}^{G}\hat{\boldsymbol{\beta}}_{(-g)}$。

MacKinnon et al. (2022b) 证明了 $\widehat{\mathbf{V}}_{\hat{\boldsymbol{\beta}}_{ols}}^{\text{CV3}}$ 具有如下等价的表达形式：

$$\widehat{\mathbf{V}}_{\hat{\boldsymbol{\beta}}_{ols}}^{\text{CV3}} = \frac{G-1}{G}\sum_{g=1}^{G}\left(\hat{\boldsymbol{\beta}}_{(-g)}-\hat{\boldsymbol{\beta}}_{ols}\right)\left(\hat{\boldsymbol{\beta}}_{(-g)}-\hat{\boldsymbol{\beta}}_{ols}\right)'.$$

因此，$\widehat{\mathbf{V}}_{\hat{\boldsymbol{\beta}}_{ols}}^{\text{CV3}}$ 实际上是一种 Jackknife 估计量。

例 3.6 Duflo et al. (2011) 利用 5795 名就读于 121 所肯尼亚小学的在校生的随机化实验数据研究按能力分班教学对学生测试成绩的影响。本例使用的数据可以从 Hansen (2022a) 的配套网站 `https://www.ssc.wisc.edu/~bhansen/econometrics/`下载，数据文件为“DDK2011.dta”。我们将学生的测试成绩（totalscore）对虚拟变量 tracking 进行回归，tracking 取 1 意味着学生就读的小学被随机分配到实验组，即，该学校需要按能力分班教学；否则，tracking 取 0；聚类变量为 schoolid，用以区分不同的小学；此外，回归时我们控制了性别（girl）和测试时的年龄（agetest）。在如下的 Stata 命令中，我们首先使用 summclust 命令得到基于式 (3.18) 的聚类稳健标准误；然后，我们分别使用 cluster(schoolid) vce(jackknife) 和 jackknife, cluster(schoolid) 得到聚类稳健 Jackknife 标准误。根据前文的分析，三个命令得到的标准误是相同的，以 tracking 的标准误为例，三个命令

得到的结果都是 0.078294。

```
. use DDK2011.dta, replace
. egen testscore = std(totalscore) //对测试成绩进行标准化处理
. summclust tracking, yvar(testscore) xvar(girl agetest)
    cluster(schoolid)
. reg testscore tracking girl agetest, cluster(schoolid)
    vce(jackknife,nodots)
. jackknife, cluster(schoolid) nodots: reg testscore tracking girl
    agetest
```

非参数 Bootstrap

Bootstrap 方法最早由 Efron (1979) 提出，此方法可以一致地估计计量模型中的估计量和检验统计量的渐近分布，因此，在计量经济学中具有广泛的应用。本质上，Bootstrap 分布的取得依赖于对原始样本数据进行的有放回[6]重抽样。MacKinnon (2006) 和 Horowitz (2019) 对此方法在计量领域的使用做了较完整地介绍。

我们首先讨论 Bootstrap 方差和标准误。令 B 表示 Bootstrap 重抽样的次数；$\hat{\boldsymbol{\theta}}$ 是基于原始样本数据得到的向量值参数 $\boldsymbol{\theta}$ 的估计量；$\hat{\boldsymbol{\theta}}^*(b)$ 表示利用第 b 次重抽样的 Bootstrap 样本数据计算的 $\boldsymbol{\theta}$ 的估计量。$\hat{\boldsymbol{\theta}}$ 的方差的 Bootstrap 估计量为：

$$\widehat{\mathbf{V}}_{\hat{\boldsymbol{\theta}}}^{\text{boot}} = \frac{1}{B-1}\sum_{b=1}^{B}\left(\hat{\boldsymbol{\theta}}^*(b)-\bar{\boldsymbol{\theta}}^*\right)\left(\hat{\boldsymbol{\theta}}^*(b)-\bar{\boldsymbol{\theta}}^*\right)', \tag{3.22}$$

其中，

$$\bar{\boldsymbol{\theta}}^* = \frac{1}{B}\sum_{b=1}^{B}\hat{\boldsymbol{\theta}}^*(b).$$

$\hat{\boldsymbol{\theta}}$ 中第 k 个元素的 Bootstrap 标准误，记为 $s_{\hat{\theta}_k}^{\text{boot}}$，是 $\widehat{\mathbf{V}}_{\hat{\boldsymbol{\theta}}}^{\text{boot}}$ 的对角线上第 k 个元素的平方根。

Bootstrap 标准误可以用来构造置信区间，θ_k 的置信水平为 $1-\alpha$ 的正态近似 Bootstrap 置信区间定义为：

$$C^{\text{nb}} = \left[\hat{\theta}_k - z_{1-\frac{\alpha}{2}}\cdot s_{\hat{\theta}_k}^{\text{boot}},\quad \hat{\theta}_k + z_{1-\frac{\alpha}{2}}\cdot s_{\hat{\theta}_k}^{\text{boot}}\right],$$

其中，$z_{1-\frac{\alpha}{2}}$ 是标准正态分布的 $1-\alpha/2$ 分位数。当使用 Bootstrap 标准误时，Stata 默认汇报的置信区间便是 C^{nb}。由于正态近似 Bootstrap 置信区间的构造依赖于 t 统计量的极限分布，在有限样本情况下，t 统计量的精确分布可能并非是标准正态分布。因此，Bootstrap 标准误更多地是被用来测度系数估计的精确程度，很少应用于构造置信区间。在 Stata 中，回归系数估计的 Bootstrap 标准误可以通过 vce(bootstrap, reps(#)) 选项得到，其中，# 表示 Bootstrap 重抽样的次数；也可以通过命令 bootstrap 得到。例如：

[6] 若是无放回重抽样，得到的 Bootstrap 样本将与原始样本一致。

```
//接例3.6的程序
. reg testscore tracking girl agetest, vce(bootstrap, reps(500))
. bootstrap, reps(500): reg testscore tracking girl agetest
```

虽然 Bootstrap 标准误不适合用于构造置信区间，但是，Bootstrap 方法依然可以用于置信区间的构造。一种常用且简单的构造方法被称为百分位数区间（percentile interval），此方法利用了 Bootstrap 分布的样本分位数。具体地，令 $\hat{\theta}$ 是标量参数 θ 的估计量，经过 B 次 Bootstrap 重抽样，我们得到一个来自 Bootstrap 估计量 $\hat{\theta}^*$ 的分布[7]的随机样本 $\{\hat{\theta}_1^*, \hat{\theta}_2^*, \ldots, \hat{\theta}_B^*\}$。对任意的 $0 < \alpha < 1$，我们可以计算这些 Bootstrap 估计量的 α 经验分位数 q_α^*，即，q_α^* 满足 $\{\hat{\theta}_1^*, \hat{\theta}_2^*, \ldots, \hat{\theta}_B^*\}$ 中不大于 q_α^* 的估计量的个数为 $n\alpha$。[8] 根据次序统计量的知识，q_α^* 的常用计算方式是取 $\{\hat{\theta}_b^*\}$ 的第 $n\alpha$ 个次序统计量。于是，置信水平为 $1-\alpha$ 的 Bootstrap 百分位数区间为：

$$C^{\mathrm{pc}} = \left[q_{\alpha/2}^*, \quad q_{1-\alpha/2}^*\right]. \tag{3.23}$$

例如，若 $B = 1000$，$\alpha = 0.05$，则 $C^{\mathrm{pc}} = [\hat{\theta}_{(25)}^*, \hat{\theta}_{(975)}^*]$。

在 Stata 中，C^{pc} 可以通过在汇报 Bootstrap 标准误的回归命令之后使用 estat bootstrap, percentile 或 estat bootstrap, all 得到。例如：

```
//接例3.6的程序
. reg testscore tracking girl agetest, vce(bootstrap, reps(500))
. estat bootstrap, percentile
. estat bootstrap, all
```

注 理论上讲，Bootstrap 实际上是估计量 $\hat{\theta}$ 的抽样分布的一种估计。具体地，令 F 表示原始样本中的一个观测服从的联合分布函数。再令 $G_n(u, F)$ 表示 $\hat{\theta}$ 的分布函数，即，

$$G_n(u, F) = \mathbb{P}\left[\hat{\theta} \leqslant u \mid F\right].$$

这里为了讨论问题的方便，假设 θ 是标量。在计算 $\hat{\theta}$ 的标准误和构造置信区间时，我们需要知道关于 G_n 的信息。通常情况下，F 是未知的，并且，G_n 的函数形式也是未知的。为了解决这两个问题，Bootstrap 首先利用经验分布函数 F_n 估计 F，将 G_n 中的 F 替换为 F_n 得到：

$$G_n^*(u) = G_n(u, F_n);$$

Bootstrap 然后利用统计模拟的方式估计 G_n^*。注意到从分布 F_n 随机抽样等价于对原始样本数据进行有放回的抽样。于是，对经过 n 次有放回抽样得到的一个 Bootstrap 样本使用估计量 $\hat{\theta}$ 的表达式，我们可以得到服从 $G_n^*(u)$ 分布的一个随机变量。将此过程大量重复 B 次，便可得到服从 $G_n^*(u)$ 分布的一个样本量为 B 的随机样本。基于此样本，我们可以估计 G_n^* 的各种描述统计量。

[7] Bootstrap 估计量 $\hat{\theta}^*$ 是基于 Bootstrap 样本按照 $\hat{\theta}$ 的函数形式计算的估计量，它的分布依赖于样本量 n 和原始样本数据的经验分布函数 F_n。

[8] 假设 $n\alpha$ 为正整数。

明确 G_n 和 G_n^* 的联系和区别对理解后续的 Bootstrap 假设检验至关重要。G_n 是估计量 $\hat{\theta}$ 的分布，计算 $\hat{\theta}$ 的样本是从总体分布 F 中随机抽样得到的；G_n^* 是 Bootstrap 估计量 $\hat{\theta}^*$ 的分布，计算 $\hat{\theta}^*$ 的样本是从经验分布 F_n 中随机抽样得到的。这里我们要区分两个“总体”：生成原始样本数据的“抽样总体”和生成 Bootstrap 样本的“Bootstrap 总体”。“抽样总体”的总体分布为 F，在此总体里参数的真实值为 θ；“Bootstrap 总体”的总体分布为 F_n，当从“Bootstrap 总体”进行抽样时，我们将 F_n 视为真实分布。在“Bootstrap 总体”中，参数的真实值为 $\hat{\theta}$。

根据上述分析，我们概括出如下非参数 Bootstrap 算法，并说明此算法的每一步与对 F_n 和 G_n^* 进行抽样之间的关系：

1. 对样本量为 n 的原始样本数据 $\{(y_1, \mathbf{x}_1), (y_2, \mathbf{x}_2), \ldots, (y_n, \mathbf{x}_n)\}$ 进行 n 次有放回的抽样得到 Bootstrap 样本 $\{(y_1^*, \mathbf{x}_1^*), (y_2^*, \mathbf{x}_2^*), \ldots, (y_n^*, \mathbf{x}_n^*)\}$。对原始样本数据做有放回的抽样等价于从经验分布 F_n 中随机抽样。
2. 基于 Bootstrap 样本 $\{(y_1^*, \mathbf{x}_1^*), (y_2^*, \mathbf{x}_2^*), \ldots, (y_n^*, \mathbf{x}_n^*)\}$，按照计算 $\hat{\theta}$ 的方式得到 Bootstrap 估计量 $\hat{\theta}^*$。由于 Bootstrap 样本是从 F_n 中随机抽样得到，因此，$\hat{\theta}^*$ 服从 G_n^* 分布，$\hat{\theta}^*$ 可以看作是从 G_n^* 分布中随机抽取的一个样本观测。
3. 将步骤 1-2 重复 B 次，得到一组 Bootstrap 估计量 $\{\hat{\theta}_1^*, \hat{\theta}_2^*, \ldots, \hat{\theta}_B^*\}$，这组随机变量可以看作是从 G_n^* 分布中随机抽取的一个样本。
4. 利用样本 $\{\hat{\theta}_1^*, \hat{\theta}_2^*, \ldots, \hat{\theta}_B^*\}$ 计算 Bootstrap 方差、标准误以及分位数。这实际上是在估计分布 G_n^* 的方差、标准差和分位数。

Bootstrap 方差估计的一致性 *

本小节讨论 Bootstrap 方差估计的渐近性质。为此，我们要适度放大 $\hat{\boldsymbol{\theta}}^* - \hat{\boldsymbol{\theta}}$，否则其方差可能退化。[9] 假设存在序列 a_n 使得：

$$\boldsymbol{z}_n = a_n(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \underset{d}{\to} \boldsymbol{\xi}, \tag{3.24}$$

$$\boldsymbol{z}_n^* = a_n(\hat{\boldsymbol{\theta}}^* - \hat{\boldsymbol{\theta}}) \underset{d^*}{\to} \boldsymbol{\xi}, \tag{3.25}$$

其中，$\boldsymbol{\xi}$ 是某一极限分布。上述假设要求：存在一种放大的方式 a_n，使得 $\hat{\boldsymbol{\theta}}$ 和 $\hat{\boldsymbol{\theta}}^*$ 的渐近分布相同。由定理 B.6 和定理 B.11，此假设对于大多数计量经济学中的估计量都是成立的。$\boldsymbol{z}_n$ 方差的 Bootstrap 估计量：

$$\widehat{\mathbf{V}}_{\hat{\boldsymbol{\theta}}}^{\text{boot,B}} = \frac{1}{B-1} \sum_{b=1}^{B} \left(\boldsymbol{z}_n^*(b) - \bar{\boldsymbol{z}}_n^*\right) \left(\boldsymbol{z}_n^*(b) - \bar{\boldsymbol{z}}_n^*\right)',$$

其中，

$$\bar{\boldsymbol{z}}_n^* = \frac{1}{B} \sum_{b=1}^{B} \boldsymbol{z}_n^*(b).$$

因为 $\boldsymbol{z}_n^*$ 依 Bootstrap 分布收敛到与 $\boldsymbol{z}_n$ 相同的渐近分布，一个合理的猜测是 $\boldsymbol{z}_n^*$ 的样本方差收敛到 $\boldsymbol{\xi}$ 的方差。然而，依分布收敛并不是依矩收敛（convergence in moments）的充

[9] 例如，在分析最小二乘估计量的大样本性质时，我们考虑的是 $\sqrt{n}(\hat{\boldsymbol{\beta}}_{ols} - \boldsymbol{\beta})$ 的渐近分布，此渐近分布的方差即为 $\hat{\boldsymbol{\beta}}_{ols}$ 的渐近方差。

分条件。为使 $\boldsymbol{z}_n^*$ 的样本方差收敛到 $\boldsymbol{\xi}$ 的方差，需要 $\boldsymbol{z}_n^*$ 一致平方可积。于是，我们有如下定理：

定理 3.7. Bootstrap 方差估计的一致性

若存在序列 a_n 使式 (3.24) 和 (3.25) 成立，并且 $\|\boldsymbol{z}_n^*\|^2$ 一致可积，则有，

$$\widehat{\mathbf{V}}_{\hat{\boldsymbol{\theta}}}^{\text{boot,B}} \underset{p^*}{\to} \widehat{\mathbf{V}}_{\hat{\boldsymbol{\theta}}}^{\text{boot}} \equiv \text{Var}[\boldsymbol{z}_n^*], \quad B \to \infty,$$

同时，随着 $n \to \infty$，

$$\widehat{\mathbf{V}}_{\hat{\boldsymbol{\theta}}}^{\text{boot}} \underset{p^*}{\to} \mathbf{V}_{\boldsymbol{\theta}} \equiv \text{Var}[\boldsymbol{\xi}].$$

♡

在计量经济学中，诸如样本方差和最小二乘估计量等常见的估计量都是一类更一般的估计量的特例，这类估计量被称为平滑函数模型（smooth function model）估计量。具体地，令 $\boldsymbol{z} \in \mathbb{R}^m$ 为 $m \times 1$ 维随机向量，$\boldsymbol{h}(\boldsymbol{z}) : \mathbb{R}^m \to \mathbb{R}^k$，$\boldsymbol{\mu} = \mathbb{E}[\boldsymbol{h}(\boldsymbol{z})]$，平滑函数模型为 $\boldsymbol{\theta} = \boldsymbol{g}(\boldsymbol{\mu})$，其中，$\boldsymbol{g}(\boldsymbol{\mu})$ 为满足某些特定条件的平滑函数。给定简单随机样本 $\{\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n\}$，参数 $\boldsymbol{\theta}$ 的平滑函数模型估计量为 $\hat{\boldsymbol{\theta}} = \boldsymbol{g}(\hat{\boldsymbol{\mu}})$，其中，$\hat{\boldsymbol{\mu}} = n^{-1}\sum_{i=1}^n \boldsymbol{h}(\boldsymbol{z}_i)$。对于平滑函数模型，我们有如下定理：

定理 3.8. 平滑函数模型的 Bootstrap 方差估计

假设 $\boldsymbol{z}_i$ 是 i.i.d. 的，$\boldsymbol{\mu} = \mathbb{E}[\boldsymbol{h}(\boldsymbol{z})]$，$\boldsymbol{\theta} = \boldsymbol{g}(\boldsymbol{\mu})$，$\mathbb{E}\|\boldsymbol{h}(\boldsymbol{z})\|^2 < \infty$，$\mathbf{G}(\boldsymbol{x}) = \dfrac{\partial}{\partial \boldsymbol{x}}\boldsymbol{g}(\boldsymbol{x})'$ 在 $\boldsymbol{\mu}$ 的某一邻域中连续。令 $\hat{\boldsymbol{\theta}} = \boldsymbol{g}(\hat{\boldsymbol{\mu}})$ 和 $\hat{\boldsymbol{\theta}}^* = \boldsymbol{g}(\hat{\boldsymbol{\mu}}^*)$，其中，$\hat{\boldsymbol{\mu}} = n^{-1}\sum_{i=1}^n \boldsymbol{h}(\boldsymbol{z}_i)$，$\hat{\boldsymbol{\mu}}^* = n^{-1}\sum_{i=1}^n \boldsymbol{h}(\boldsymbol{z}_i^*)$。如果存在 $p \geqslant 1$ 使得 $\boldsymbol{g}(\boldsymbol{x})$ 的 p 阶导函数有界，则有，$\boldsymbol{z}_n^* = \sqrt{n}(\hat{\boldsymbol{\theta}}^* - \hat{\boldsymbol{\theta}})$ 一致平方可积，由定理 3.7，Bootstrap 方差估计是一致的。

♡

此定理的证明请参考 Hansen (2022a) 的第 10.31 节。

Bootstrap 方差的截尾估计量 *

对于具有有界 p 阶导数的平滑函数，定理 3.8 表明其方差的 Bootstrap 估计量是一致的。虽然此结论适用于计量经济学中大多数常见的估计量，但是，对于一些重要的非线性形式的函数 $\boldsymbol{g}(\cdot)$，此结论并不成立。例如，$\boldsymbol{z}_i = (z_{i1}, z_{i2})$，$\theta = \mu_1/\mu_2$，其中，$\mu_1 = \mathbb{E}[z_{i1}]$ 和 $\mu_2 = \mathbb{E}[z_{i2}]$。导致定理 3.8 的结论不成立的原因是 $\hat{\theta}$ 不存在有限的渐近方差。

为了缓解渐近方差无限对方差估计带来的影响，可以考虑使用截尾的 Bootstrap 方差估计量。具体地，令 $\tau_n \to \infty$ 是取值为正的截尾数列，定义截尾变量：

$$\boldsymbol{z}_n^{**} = \boldsymbol{z}_n^* \cdot \mathbf{1}\{\|\boldsymbol{z}_n^*\| \leqslant \tau_n\}.$$

$\hat{\theta}$ 的渐近方差的截尾 Bootstrap 估计量为：

$$\widehat{\mathbf{V}}_{\hat{\boldsymbol{\theta}}}^{\text{boot,B},\tau} = \frac{1}{B-1}\sum_{b=1}^B (\boldsymbol{z}_n^{**}(b) - \bar{\boldsymbol{z}}_n^{**})(\boldsymbol{z}_n^{**}(b) - \bar{\boldsymbol{z}}_n^{**})',$$

其中，

$$\bar{\boldsymbol{z}}_n^{**} = \frac{1}{B}\sum_{b=1}^B \boldsymbol{z}_n^{**}(b).$$

关于截尾估计量，我们有如下结论：

定理 3.9. 截尾 Bootstrap 方差估计的一致性

若存在序列 a_n 使式 (3.24) 和 (3.25) 成立，则有，

$$\widehat{\mathbf{V}}_{\hat{\boldsymbol{\theta}}}^{\text{boot,B},\tau} \underset{p^*}{\to} \widehat{\mathbf{V}}_{\hat{\boldsymbol{\theta}}}^{\text{boot},\tau} \equiv \text{Var}[\boldsymbol{z}_n^{**}], \quad B \to \infty,$$

同时，若定理 3.8 中除去 p 阶导函数有界的其余假设都成立，则有，

$$\widehat{\mathbf{V}}_{\hat{\boldsymbol{\theta}}}^{\text{boot},\tau} \underset{p^*}{\to} \mathbf{V}_{\boldsymbol{\theta}} \equiv \text{Var}[\boldsymbol{\xi}], \quad n \to \infty.$$

♡

定理 3.9 表明截尾 Bootstrap 方差估计量是平滑函数模型渐近方差的一致估计量。这一结论的成立关键在于估计渐近方差时要对 $\boldsymbol{z}_n^*$ 进行截尾。然而，Stata 在计算 Bootstrap 标准误时使用的是非截尾的估计量 (3.22)，因此，Stata 汇报的标准误可能是不一致的。式 (3.22) 的非截尾 Bootstrap 方差估计量只适用于定理 3.8 考虑的 p 阶导数有界的平滑函数模型以及其余满足 $\boldsymbol{z}_n^*$ 一致平方可积的情形。定理 3.9 的证明请参考 Hansen (2022a) 的第 10.31 节。

例 3.7 （Hansen (2022a) 的第 10.15 节）本例使用 2009 年 3 月的 Current Population Survey (CPS) 数据，原始数据的下载地址为 `https://www.ssc.wisc.edu/~bhansen/econometrics/`，数据文件为“cps09mar.dta”。我们从中提取了 982 名已婚黑人妇女的年龄、收入和受教育情况等特征，处理后的数据可以从本书网站下载，其文件名为“cps09mar_subset.dta”。考虑如下回归：

$$\log(wage) = \beta_0 + \beta_1 \cdot education + \beta_2 \cdot experience + \beta_3 \cdot experience^2/100 + \varepsilon,$$

其中，

$$experience = age - education - 6.$$

通过 Stata 命令

```
. use cps09mar_subset.dta, replace
. gen experience = age - education - 6
. gen exp2 = experience^2/100
. gen lw = log(earnings)
. reg lw education experience exp2, vce(hc2)
. mat list e(V)     //返回系数估计量的协方差矩阵
```

可以得到如下回归结果（括号里为标准误）：

$$\widehat{\log(wage)} = \underset{(0.159)}{8.518} + \underset{(0.008)}{0.123}\, education + \underset{(0.006)}{0.017}\, experience - \underset{(0.012)}{0.025}\, experience^2/100.$$

系数估计量的协方差矩阵为：

$$\widehat{\mathbf{V}}^{\mathrm{HC2}}_{\hat{\beta}_{ols}} = \begin{pmatrix} 0.669 & 0.130 & -0.136 & -11.63 \\ 0.130 & 0.393 & -0.733 & -6.303 \\ -0.136 & -0.733 & 1.478 & 9.400 \\ -11.63 & -6.303 & 9.400 & 254.2 \end{pmatrix} \times 10^{-4}.$$

接下来我们考虑最大化期望对数收入的 experience 水平 θ，容易验证 $\theta = -50\beta_2/\beta_3$。利用如下 Stata 命令可以计算 θ 的估计值：

```
. nlcom -50*_b[experience]/_b[exp2] //计算结果为34.9
```

$\hat{\theta}$ 基于渐近理论的标准误可以由 Delta 方法 B.5 求得，具体地，

$$s(\hat{\theta}) = \sqrt{\begin{pmatrix} -50/\hat{\beta}_3 & 50\hat{\beta}_2/\hat{\beta}_3^2 \end{pmatrix} \begin{pmatrix} 0.393 & -0.733 \\ -0.733 & 1.478 \end{pmatrix} \begin{pmatrix} -50/\hat{\beta}_3 \\ 50\hat{\beta}_2/\hat{\beta}_3^2 \end{pmatrix} \times 10^{-4}} \approx 5.64.$$

如下 Stata 命令可以汇报 θ 的 Jackknife 标准误：

```
. jackknife (-50*_b[experience]/_b[exp2]), nodots: ///
    reg lw education experience exp2
```

从输出结果看，$\hat{\theta}$ 的 Jackknife 标准误为 5.98，与渐近标准误比较接近。$\hat{\theta}$ 的 Bootstrap 标准误（非截尾）可以通过如下 Stata 命令得到：[10]

```
. bootstrap (-50*_b[experience]/_b[exp2]), reps(10000) ///
    nodots seed(123): reg lw education experience exp2
```

在上面的命令里，我们取 $B = 10000$，并将随机数的种子设置为 123。经过“漫长”的等待[11]，Stata 输出的 Bootstrap 标准误为 6873.98！如下 Stata 命令将随机数的种子设置为 19810814，重新计算 Bootstrap 标准误，其输出的 Bootstrap 标准误为 647.71！

```
. bootstrap (-50*_b[experience]/_b[exp2]), reps(10000) ///
    nodots seed(19810814): reg lw education experience exp2
```

从上述结果看，Bootstrap 标准误与渐近标准误之间相差悬殊，不同次 Bootstrap 得到的标准误之间也存在较大的差异，这是一致平方可积条件不成立的信号，非截尾的 Bootstrap 标准误在此种情况下是不可靠的。

最后，我们将 τ_n 取为 535（略大于渐近标准误的 $3\sqrt{n}$ 倍），计算截尾的 Bootstrap 标准误。因为 Stata 不汇报截尾的 Bootstrap 标准误，我们使用 R 语言自行编写截尾的 Bootstrap 函数。

```
#数据预处理
```

[10] 做好准备，感受一下 Stata 的计算速度有多慢！

[11] 运行时间大概为 9 分 15 秒。

```r
library(tidyverse)
#读取Stata数据文件
cps09mar_subset <- haven::read_dta("~/cps09mar_subset.dta")
reg_data <- cps09mar_subset %>%
  #利用mutate命令创建新变量experience, lw和exp2
  dplyr::mutate(experience=age-education-6,lw=log(earnings)) %>%
  dplyr::mutate(exp2=experience^2/100) %>%
  #利用select命令筛选出lw, education, experience和exp2四个变量
  dplyr::select(c(lw,education,experience,exp2))

#截尾Bootstrap函数
"trimBootstrap" <- function(sample_data, tau=20, nter=1000){
  #sample_data: 原始样本数据集
  #tau: trimming number
  #nter: B

  #利用原始数据进行回归
  fit <- lm(lw~education+experience+exp2,data=sample_data)
  beta2 <- coefficients(fit)["experience"]
  beta3 <- coefficients(fit)["exp2"]
  hat_theta <- as.numeric(-50*beta2/beta3)  #计算参数theta的估计值
  n <- dim(sample_data)[1]                  #获取样本量
  theta_star=NULL                           #定义空串theta_star
  #B次重抽样并利用每次重抽样得到的Bootstrap样本计算theta的估计值
  for (b in 1:nter){
    #对不大于n的正整数进行有放回的随机抽样
    index <- sample(n,replace = TRUE)
    #根据抽样结果生成Bootstrap样本
    current_data <- sample_data[index,]
    #利用Bootstrap样本进行回归并计算参数theta的Bootstrap估计值
    current_fit <- lm(lw~education+experience+exp2,data=current_data)
    current_beta2 <- coefficients(current_fit)["experience"]
    current_beta3 <- coefficients(current_fit)["exp2"]
    #将当前theta的Bootstrap估计值与之前循环得到的所有估计值进行合并
    theta_star = c(theta_star,
        as.numeric(-50*current_beta2/current_beta3))
  }
  #计算式(3.25)定义的变量
```

```
  z_star = sqrt(n)*(theta_star - hat_theta)
  #对z_star进行截尾
  z_star2 = z_star*(abs(z_star) <= tau)
  #计算z_star2的样本方差，即为截尾的Bootstrap渐近方差估计
  trimmed_boot_var = var(z_star2)
  #计算截尾的Bootstrap标准误
  boot_sd = sqrt(trimmed_boot_var/n)
  #返回截尾数据和截尾的Bootstrap标准误
  trimBootstrap <- list(bootvalue=z_star2,trimmed_sd=boot_sd)
}

#调用trimBootstrap函数计算截尾Bootstrap标准误
sd <- trimBootstrap(reg_data, tau=535, nter=10000)$trimmed_sd
```

运行上述 R 程序（运行时间约为 8 秒）得到的截尾 Bootstrap 标准误近似为 5，并且，不同次 Bootstrap 计算的截尾标准误都十分接近。

偏误修正的百分位数区间 *

置信水平为 $1-\alpha$ 的 Bootstrap 百分位数区间 (3.23) 具有如下性质：

定理 3.10. C^{pc} 的渐近覆盖率

若存在序列 a_n 使得：

$$a_n(\hat{\theta}-\theta) \underset{d}{\to} \xi, \tag{3.26}$$

$$a_n(\hat{\theta}^*-\hat{\theta}) \underset{d^*}{\to} \xi, \tag{3.27}$$

其中，ξ 是关于 0 对称的连续分布，则有：当 $n\to\infty$ 时，$\mathbb{P}\left[\theta\in C^{\text{pc}}\right]\to 1-\alpha$。

证明 由式 (3.27)，$a_n(\hat{\theta}^*-\hat{\theta})$ 的 α 分位数 $a_n(q_\alpha^*-\hat{\theta})$ 依概率收敛到 ξ 的 α 分位数，记为 $\bar{q}_\alpha$，即，

$$a_n(q_\alpha^*-\hat{\theta}) \underset{p}{\to} \bar{q}_\alpha.$$

令 $H(x)=\mathbb{P}[\xi\leqslant x]$ 是 ξ 的分布函数，由对称性，$H(-x)=1-H(x)$。于是，置信水平为 $1-\alpha$ 的 Bootstrap 百分位数区间 (3.23) 覆盖真实参数的概率为：

$$\begin{aligned}
\mathbb{P}\left[\theta\in C^{\text{pc}}\right] &= \mathbb{P}\left[q_{\alpha/2}^*\leqslant\theta\leqslant q_{1-\alpha/2}^*\right] \\
&= \mathbb{P}\left[-a_n\left(q_{\alpha/2}^*-\widehat{\theta}\right)\geqslant a_n(\widehat{\theta}-\theta)\geqslant -a_n\left(q_{1-\alpha/2}^*-\widehat{\theta}\right)\right] \\
&\to \mathbb{P}\left[-\bar{q}_{\alpha/2}\geqslant\xi\geqslant-\bar{q}_{1-\alpha/2}\right] \\
&= H\left(-\bar{q}_{\alpha/2}\right)-H\left(-\bar{q}_{1-\alpha/2}\right) \\
&= H\left(\bar{q}_{1-\alpha/2}\right)-H\left(\bar{q}_{\alpha/2}\right) \\
&= 1-\alpha.
\end{aligned}$$

□

当极限分布 ξ 对称时，定理 3.10 表明百分位数区间是渐近有效的。由定理 B.6 和定理 B.11，平滑函数模型满足 (3.26) 和 (3.27)，因此定理 3.10 的结论适用于计量经济学中的常用估计量。然而，在小样本下，估计量的真实分布可能是非对称的，百分位数区间在小样本下的真实覆盖率可能与名义置信水平存在较大差异。即便如此，小样本下使用百分位数区间依然具有一定的合理性。假设存在严格递增的未知函数 $\psi(\theta)$ 使得 $\psi(\hat{\theta})-\psi(\theta)$ 具有关于 0 对称的枢轴（pivotal）分布[12] $H(u)$。例如，若 $\hat{\theta}=\exp(\hat{\mu})$，$\hat{\mu}\sim\mathcal{N}(\mu,\sigma^2)$，可以取 $\psi(\theta)=\log(\theta)/\sigma$。注意到 H 是枢轴分布，因此 $\psi(\hat{\theta}^*)-\psi(\hat{\theta})$ 的分布也为 $H(u)$。令 $\bar{q}_\alpha$ 是分布 H 的 α 分位数。因为 q_α^* 是 $\hat{\theta}^*$ 的 α 分位数，并且 $\psi(\hat{\theta}^*)-\psi(\hat{\theta})$ 是单调变换，因此 $\bar{q}_\alpha+\psi(\hat{\theta})=\psi(q_\alpha^*)$。于是，我们有：

$$
\begin{aligned}
\mathbb{P}\left[\theta\in C^{\mathrm{pc}}\right] &= \mathbb{P}\left[q_{\alpha/2}^*\leqslant\theta\leqslant q_{1-\alpha/2}^*\right]\\
&= \mathbb{P}\left[\psi\left(q_{\alpha/2}^*\right)\leqslant\psi(\theta)\leqslant\psi\left(q_{1-\alpha/2}^*\right)\right]\\
&= \mathbb{P}\left[\psi(\hat{\theta})-\psi\left(q_{\alpha/2}^*\right)\geqslant\psi(\hat{\theta})-\psi(\theta)\geqslant\psi(\hat{\theta})-\psi\left(q_{1-\alpha/2}^*\right)\right]\\
&= \mathbb{P}\left[-\bar{q}_{\alpha/2}\geqslant\psi(\hat{\theta})-\psi(\theta)\geqslant-\bar{q}_{1-\alpha/2}\right]\\
&= H\left(-\bar{q}_{\alpha/2}\right)-H\left(-\bar{q}_{1-\alpha/2}\right)\\
&= H\left(\bar{q}_{1-\alpha/2}\right)-H\left(\bar{q}_{\alpha/2}\right)\\
&= 1-\alpha.
\end{aligned}
$$

上述推导的结果表明：为了使百分位数区间以正确的概率覆盖 θ，$\hat{\theta}$ 的分布可以是非对称的。前提是存在单调变换 $\psi(\cdot)$ 使得 $\psi(\hat{\theta})-\psi(\theta)$ 具有关于 0 对称的枢轴分布。在这种情况下，百分位数区间的覆盖率与名义的置信水平是一致的。例如，假设我们感兴趣的参数是 $\theta=\exp(\mu)$，其中，$\mu=\mathbb{E}(\mathrm{y})$，并且 y 具有关于 μ 对称的枢轴分布。令 $\{y_1,\ldots,y_n\}$ 是 y 的一组观测值，其样本均值为 $\bar{y}$。虽然 θ 的估计量 $\hat{\theta}=\exp(\bar{y})$ 的分布是非对称的，但是 θ 的百分位数区间依然具有正确的覆盖率，这是因为 $\log(\hat{\theta})$ 具有对称的枢轴分布。

上述对百分位数方法合理性的辩护依赖于 $\psi(\hat{\theta})-\psi(\theta)$ 具有关于 0 对称的枢轴分布，其必要条件是 $\mathbb{E}[\psi(\hat{\theta})]=\psi(\theta)$。若 $\mathbb{E}[\psi(\hat{\theta})]\neq\psi(\theta)$，百分位数区间的覆盖率通常会偏离名义水平。为了解决 $\psi(\hat{\theta})$ 有偏的问题，Efron (1982) 提出了偏误修正（bias-corrected, BC）的百分位数区间。假设存在未知的严格递增变换 $\psi(\theta)$ 和未知的常数 z_0 使得：

$$
z=\psi(\hat{\theta})-\psi(\theta)+z_0\sim\mathcal{N}(0,1). \tag{3.28}
$$

令 $\Phi(x)$ 为标准正态分布的累积分布函数，则 $\Phi^{-1}(p)$ 为标准正态分布的 p 分位数。BC 区间可以利用 Bootstrap 估计量 $\hat{\theta}_b^*$ 和 Bootstrap 分位数 q_α^* 构造。具体地，取

$$
p^*=\frac{1}{B}\sum_{b=1}^{B}\mathbf{1}\left\{\hat{\theta}_b^*\leqslant\hat{\theta}\right\},\quad z_0^*=\Phi^{-1}(p^*). \tag{3.29}
$$

p^* 是中位数偏误[13]（median bias）的度量，z_0^* 将此偏误映射为标准正态的分位数。（若 $\hat{\theta}$

[12] 分布与参数 θ 无关。

[13] 偏误定义为估计量的期望与参数真实值之间的差；中位数偏误是估计量的分布的中位数与参数真实值之间的差。如果估计量是中位数无偏的，则它的单射变换也是中位数无偏的。

的中位数偏误为 0，则 $p^* = 0.5$，$z_0^* = 0$；若 $\hat{\theta}$ 上偏（upwards biased），则 $p^* < 0.5$，$z_0^* < 0$；若 $\hat{\theta}$ 下偏（dowward biased），则 $p^* > 0.5$，$z_0^* > 0$。）对任意的 $\alpha \in (0,1)$，定义

$$x(\alpha) = \Phi(z_\alpha + 2z_0^*),$$

其中，$z_\alpha = \Phi^{-1}(\alpha)$。（若 $z_0^* = 0$，则 $x(\alpha) = \alpha$；若 $z_0^* > 0$，则 $x(\alpha) > \alpha$；若 $z_0^* < 0$，则 $x(\alpha) < \alpha$。）BC 百分位数区间定义为：

$$C^{\text{bc}} = \left[q^*_{x(\alpha/2)}, \quad q^*_{x(1-\alpha/2)}\right]. \tag{3.30}$$

如果假设 (3.28) 成立，可以证明 C^{bc} 具有正确的覆盖率。由 (3.28)，

$$\mathbb{P}\left[\psi(\hat{\theta}) - \psi(\theta) + z_0 \leqslant x\right] = \Phi(x).$$

因为 $\Phi(x)$ 是枢轴分布，我们有：

$$\mathbb{P}^*\left[\psi(\hat{\theta}^*) - \psi(\hat{\theta}) + z_0 \leqslant x\right] = \Phi(x). \tag{3.31}$$

在上式中取 $x = z_0$ 得到：

$$\mathbb{P}^*\left[\psi(\hat{\theta}^*) - \psi(\hat{\theta}) \leqslant 0\right] = \Phi(z_0),$$

进而可以推出

$$\mathbb{P}^*\left[\hat{\theta}^* \leqslant \hat{\theta}\right] = \Phi(z_0).$$

对上式两端取 $\Phi(\cdot)$ 的反函数得到：

$$z_0 = \Phi^{-1}\left(\mathbb{P}^*\left[\hat{\theta}^* \leqslant \hat{\theta}\right]\right).$$

上式即为 (3.29) 中的 z_0^* 在 $B \to \infty$ 情况下的概率极限。因此，我们可以利用 z_0^* 近似 z_0。在下面的讨论中，我们假设 z_0 已知，其取值等于 z_0^*。根据式 (3.31)，

$$\begin{aligned} x(\alpha) &= \Phi(z_\alpha + 2z_0) \\ &= \mathbb{P}^*\left[\psi(\hat{\theta}^*) - \psi(\hat{\theta}) \leqslant z_\alpha + z_0\right] \\ &= \mathbb{P}^*\left[\hat{\theta}^* \leqslant \psi^{-1}\left(\psi(\hat{\theta}) + z_0 + z_\alpha\right)\right]. \end{aligned}$$

上式说明 $q^*_{x(\alpha)} = \psi^{-1}(\psi(\hat{\theta}) + z_0 + z_\alpha)$。因此，$C^{\text{bc}}$ 可以写成：

$$C^{\text{bc}} = \left[\psi^{-1}\left(\psi(\hat{\theta}) + z_0 + z_{\alpha/2}\right), \quad \psi^{-1}\left(\psi(\hat{\theta}) + z_0 + z_{1-\alpha/2}\right)\right].$$

根据上式，C^{bc} 的覆盖率为：

$$\begin{aligned} \mathbb{P}[\theta \in C^{\text{bc}}] &= \mathbb{P}\left[\psi^{-1}\left(\psi(\hat{\theta}) + z_0 + z_{\alpha/2}\right) \leqslant \theta \leqslant \psi^{-1}\left(\psi(\hat{\theta}) + z_0 + z_{1-\alpha/2}\right)\right] \\ &= \mathbb{P}\left[\psi(\hat{\theta}) + z_0 + z_{\alpha/2} \leqslant \psi(\theta) \leqslant \psi(\hat{\theta}) + z_0 + z_{1-\alpha/2}\right] \\ &= \mathbb{P}\left[-z_{\alpha/2} \geqslant \psi(\hat{\theta}) - \psi(\theta) + z_0 \geqslant -z_{1-\alpha/2}\right] \\ &= \mathbb{P}\left[z_{1-\alpha/2} \geqslant z \geqslant z_{\alpha/2}\right] \\ &= \Phi\left(z_{1-\alpha/2}\right) - \Phi\left(z_{\alpha/2}\right) \\ &= 1 - \alpha, \end{aligned}$$

其中，第四个等号用到了假设 (3.28) 以及 $z_\alpha = -z_{1-\alpha}$。

在 Stata 中，C^{bc} 可以通过在汇报 Bootstrap 标准误的回归命令之后使用 estat boot-

strap 得到。例如：

```
//接例3.6的程序
. reg testscore tracking girl agetest, vce(bootstrap, reps(500))
. estat bootstrap
```

Bootstrap 加速偏误修正百分位数区间 *

虽然正态近似置信区间 $\hat{\theta} \pm z_{1-\alpha/2} \cdot s_{\hat{\theta}}^{\text{boot}}$ 在实证分析中被广泛使用，但是其真实的覆盖率与名义置信水平可能存在较大的差异。通常，我们有：

$$\mathbb{P}\left[\theta \leqslant \hat{\theta} + z_{\alpha} \cdot s_{\hat{\theta}}^{\text{boot}}\right] = \alpha + c_1/\sqrt{n}, \tag{3.32}$$

其中，c_1 为某一常数。式 (3.32) 定义了“一阶精度”（first-order accuracy），其收敛到名义值 α 的速度较慢。

本小节讨论具有“二阶精度”（second-order accuracy）水平的置信区间的构造方法。令 $\hat{\theta}_{\alpha}^{\text{2nd}}$ 为根据某种“二阶精度”方法估计的参数分布的 α 分位数，则“二阶精度”意味着：

$$\mathbb{P}\left[\theta \leqslant \hat{\theta}_{\alpha}^{\text{2nd}}\right] = \alpha + c_2/n, \tag{3.33}$$

其中，c_2 为某一常数。由此可见，“二阶精度”方法的收敛速度更快。

前文介绍的 Bootstrap 百分位数区间和 BC 区间都不是具有“二阶精度”的方法。BC 区间不是具有“二阶精度”的方法的原因在于假设 (3.28) 要求 $\psi(\hat{\theta})$ 的方差为常数。Efron (1987) 提出的 Bootstrap 加速的偏误修正（accelerated bias-corrected, BC$_a$）百分位数区间假设：

$$z = \frac{\psi(\hat{\theta}) - \psi(\theta)}{1 + a\psi(\theta)} + z_0 \sim \mathcal{N}(0,1), \quad 1 + a\psi(\theta) > 0. \tag{3.34}$$

在此假设下，$\psi(\hat{\theta})$ 的方差与 θ 和加速参数 a 有关。z_0 可以用 (3.29) 中的 z_0^* 代替，Efron (1987) 建议使用 $\hat{\theta}$ 的偏度的缩放（scaled）Jackknife 估计量来估计加速参数 a：

$$\hat{a} = \frac{\sum_{i=1}^{n}\left(\bar{\theta} - \hat{\theta}_{(-i)}\right)^3}{6\left[\sum_{i=1}^{n}\left(\bar{\theta} - \hat{\theta}_{(-i)}\right)^2\right]^{3/2}}, \quad \bar{\theta} = \frac{1}{n}\sum_{i=1}^{n}\hat{\theta}_{(-i)}.$$

对任意的 $\alpha \in (0,1)$，定义

$$x(\alpha) = \Phi\left(z_0^* + \frac{z_{\alpha} + z_0^*}{1 - \hat{a}\left(z_{\alpha} + z_0^*\right)}\right).$$

BC$_a$ 百分位数区间为：

$$C^{\text{bca}} = \left[q_{x(\alpha/2)}^*, \quad q_{x(1-\alpha/2)}^*\right].$$

注意到当 $\hat{a} = 0$ 时，C^{bca} 等价于 C^{bc}。C^{bc} 对 C^{pc} 的改进是修正了中位数偏误，而 C^{bca} 进一步对偏度进行了修正。此外，C^{bca} 只有在 $\hat{a}\left(z_{\alpha} + z_0^*\right) < 1$ 时才是良定义的。

接下来，我们说明假设 (3.34) 可以推出 C^{bca} 具有正确的覆盖率。在这里，我们将假设 (3.34) 中的 a 当作是已知的常数。a 未知且需要用 $\hat{a}$ 进行估计的情况请读者参考 Shao

et al. (1995) 的第四章。由 (3.34)，

$$\mathbb{P}^*\left[\frac{\psi(\hat{\theta}^*)-\psi(\hat{\theta})}{1+a\psi(\hat{\theta})}+z_0\leqslant x\right]=\Phi(x). \tag{3.35}$$

在上式中取 $x=z_0$ 得到：

$$\mathbb{P}^*\left[\frac{\psi(\hat{\theta}^*)-\psi(\hat{\theta})}{1+a\psi(\hat{\theta})}\leqslant 0\right]=\Phi(z_0).$$

根据 $\psi(\cdot)$ 严格递增且 $1+a\psi(\hat{\theta})>0$，我们有：

$$\mathbb{P}^*\left[\hat{\theta}^*\leqslant\hat{\theta}\right]=\Phi(z_0).$$

因此，

$$z_0=\Phi^{-1}\left(\mathbb{P}^*\left[\hat{\theta}^*\leqslant\hat{\theta}\right]\right).$$

上式表明 (3.29) 中的 z_0^* 依然可以用来估计 (3.34) 中的 z_0。在下面的讨论中，我们假设 z_0 已知，其取值等于 z_0^*。由 (3.35)，

$$\begin{aligned}x(\alpha)&=\mathbb{P}^*\left[\frac{\psi(\hat{\theta}^*)-\psi(\hat{\theta})}{1+a\psi(\hat{\theta})}\leqslant\frac{z_\alpha+z_0}{1-a\left(z_\alpha+z_0\right)}\right]\\&=\mathbb{P}^*\left[\hat{\theta}^*\leqslant\psi^{-1}\left(\frac{\psi(\hat{\theta})+z_\alpha+z_0}{1-a\left(z_\alpha+z_0\right)}\right)\right].\end{aligned}$$

上式说明 $q^*_{x(\alpha)}=\psi^{-1}\left(\frac{\psi(\hat{\theta})+z_\alpha+z_0}{1-a(z_\alpha+z_0)}\right)$。因此，$C^{\text{bca}}$ 可以写成：

$$C^{\text{bca}}=\left[\psi^{-1}\left(\frac{\psi(\hat{\theta})+z_{\alpha/2}+z_0}{1-a\left(z_{\alpha/2}+z_0\right)}\right),\quad\psi^{-1}\left(\frac{\psi(\hat{\theta})+z_{1-\alpha/2}+z_0}{1-a\left(z_{1-\alpha/2}+z_0\right)}\right)\right].$$

根据上式，C^{bca} 的覆盖率为：

$$\begin{aligned}\mathbb{P}\left[\theta\in C^{\text{bca}}\right]&=\mathbb{P}\left[\psi^{-1}\left(\frac{\psi(\hat{\theta})+z_{\alpha/2}+z_0}{1-a\left(z_{\alpha/2}+z_0\right)}\right)\leqslant\theta\leqslant\psi^{-1}\left(\frac{\psi(\hat{\theta})+z_{1-\alpha/2}+z_0}{1-a\left(z_{1-\alpha/2}+z_0\right)}\right)\right]\\&=\mathbb{P}\left[\frac{\psi(\hat{\theta})+z_{\alpha/2}+z_0}{1-a\left(z_{\alpha/2}+z_0\right)}\leqslant\psi(\theta)\leqslant\frac{\psi(\hat{\theta})+z_{1-\alpha/2}+z_0}{1-a\left(z_{1-\alpha/2}+z_0\right)}\right]\\&=\mathbb{P}\left[-z_{\alpha/2}\geqslant\frac{\psi(\hat{\theta})-\psi(\theta)}{1+a\psi(\theta)}+z_0\geqslant-z_{1-\alpha/2}\right]\\&=\mathbb{P}\left[z_{1-\alpha/2}\geqslant z\geqslant z_{\alpha/2}\right]\\&=1-\alpha,\end{aligned}$$

其中，第四个等号用到了假设 (3.34) 以及 $z_\alpha=-z_{1-\alpha}$。

在 Stata 中，C^{bca} 可以通过在汇报 Bootstrap 标准误并使用 bca 选项的回归命令之后使用 estat bootstrap, bca 或 estat bootstrap, all 得到。例如：

```
//接例3.6的程序
. reg testscore tracking girl agetest, vce(bootstrap, reps(500) bca
    seed(123))
. estat bootstrap, all
. estat bootstrap, bca
```

```
. bootstrap, reps(500) bca seed(123): reg testscore tracking girl
    agetest
. estat bootstrap, bca
```

除了 BC_a 百分位数区间之外，还有其他方法可以用来构造满足“二阶精度”要求的置信区间，t-百分位数区间便是其中的一种，感兴趣的读者可以参考 Hansen (2022a) 的第 10.19 节。

Bootstrap 假设检验

为了检验 $\mathbb{H}_0 : \theta = \theta_0$，备择假设为 $\mathbb{H}_1 : \theta \neq \theta_0$，通常做法是构造 t 统计量：

$$t = \frac{\hat{\theta} - \theta_0}{s(\hat{\theta})},$$

并利用标准正态的临界值进行统计推断。在小样本情况下，t 统计量的真实分布可能并非标准正态，此时可以考虑使用 Bootstrap t 检验。在 Bootstrap t 检验中，临界值是利用 Bootstrap 算法计算的。

Bootstrap t 检验的步骤如下：

1. 对样本量为 n 的原始样本数据进行 n 次有放回的抽样得到 Bootstrap 样本，利用此样本计算 Bootstrap t 统计量：
$$t^* = \frac{\hat{\theta}^* - \hat{\theta}}{s(\hat{\theta}^*)},$$
其中，$\hat{\theta}^*$ 是 θ 的 Bootstrap 估计量，$s(\hat{\theta}^*)$ 是对 Bootstrap 样本利用渐近理论计算的标准误；
2. 将步骤 1 重复 B 次，得到：$t^*(b) = (\hat{\theta}^*(b) - \hat{\theta})/s(\hat{\theta}^*(b))$，$b = 1, \ldots, B$；
3. 计算 Bootstrap p 值：
$$p^* = \frac{1}{B}\sum_{b=1}^{B} \mathbf{1}\left(|t^*(b)| > |t|\right);$$
4. 给定显著性水平 α，如果 $p^* \geqslant \alpha$，则接受 $\mathbb{H}_0$；否则，拒绝 $\mathbb{H}_0$。

当标准误没有解析表达式或不可靠时，我们可以使用 $t = \hat{\theta} - \theta_0$，相应的 Bootstrap 统计量为 $t^* = \hat{\theta}^* - \hat{\theta}$。计算 Bootstrap p 值的公式为：

$$p^* = \frac{1}{B}\sum_{b=1}^{B} \mathbf{1}\left(\left|\hat{\theta}^*(b) - \hat{\theta}\right| > \left|\hat{\theta} - \theta_0\right|\right).$$

Bootstrap t 检验适用于标量 θ，若 $\boldsymbol{\theta}$ 为向量，可以考虑使用 Bootstrap Wald 检验。Bootstrap Wald 检验的步骤如下：

1. 利用原始样本数据计算估计量 $\hat{\boldsymbol{\theta}}$ 和协方差矩阵估计量 $\widehat{\mathbf{V}}_{\hat{\boldsymbol{\theta}}}$，以及 Wald 统计量：
$$W = \left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right)' \widehat{\mathbf{V}}_{\hat{\boldsymbol{\theta}}}^{-1} \left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right);$$
2. 对原始样本数据进行 n 次有放回的抽样得到 Bootstrap 样本，利用此样本计算 Boot-

strap Wald 统计量：

$$W^* = \left(\hat{\boldsymbol{\theta}}^* - \hat{\boldsymbol{\theta}}\right)' \widehat{\mathbf{V}}_{\hat{\boldsymbol{\theta}}^*}^{-1} \left(\hat{\boldsymbol{\theta}}^* - \hat{\boldsymbol{\theta}}\right);$$

3. 将步骤 2 重复 B 次，得到：$W^*(b) = \left(\hat{\boldsymbol{\theta}}^*(b) - \hat{\boldsymbol{\theta}}\right)' \widehat{\mathbf{V}}_{\hat{\boldsymbol{\theta}}^*(b)}^{-1} \left(\hat{\boldsymbol{\theta}}^*(b) - \hat{\boldsymbol{\theta}}\right)$，$b = 1, \ldots, B$；
4. 计算 Bootstrap p 值：

$$p^* = \frac{1}{B} \sum_{b=1}^{B} \mathbf{1}\left(W^*(b) > W\right);$$

5. 给定显著性水平 α，如果 $p^* \geqslant \alpha$，则接受 $\mathbb{H}_0$；否则，拒绝 $\mathbb{H}_0$。

若可靠的协方差矩阵估计量 $\widehat{\mathbf{V}}_{\hat{\boldsymbol{\theta}}}$ 不存在，Bootstrap Wald 检验可以使用任意的正定权重矩阵代替 $\widehat{\mathbf{V}}_{\hat{\boldsymbol{\theta}}}$。

参数 Bootstrap

在计量经济学中，我们经常会设定被解释变量条件分布的参数形式。例如，在古典线性回归模型中，我们假设：

$$\mathbf{y} \mid \mathbf{X} \sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I}_n).$$

若此模型设定正确，我们可以固定解释变量 $\mathbf{x}_i$ 并从分布 $\mathcal{N}(\mathbf{x}_i'\hat{\boldsymbol{\beta}}_{ols}, s^2)$ 随机地抽取 Bootstrap 观测 y_i^*，这等价于从分布 $\mathcal{N}(0, s^2)$ 随机地抽取 e_i^* 并按如下式子计算被解释变量的 Bootstrap 观测值 y_i^*：

$$y_i^* = \mathbf{x}_i'\hat{\boldsymbol{\beta}}_{ols} + e_i^*.$$

上述产生 Bootstrap 样本的方式被称为参数 Bootstrap。若参数模型设定正确，参数 Bootstrap 的小样本性质要优于非参数 Bootstrap。

残差 Bootstrap

假设误差项服从正态分布可能并不符合某些经济或金融数据的特征，此时我们可以考虑使用残差 Bootstrap 进行统计推断。考虑如下回归模型：

$$y_i = \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i, \quad \mathbb{E}(\varepsilon_i \mid \mathbf{x}_i) = 0, \quad \varepsilon_i \sim i.i.d.\ (0, \sigma^2).$$

这里我们没有假设 ε_i 服从正态分布或其他已知分布，但是要求 ε_i 独立同分布。获取残差 Bootstrap 样本的步骤为：

1. 计算最小二乘估计量 $\hat{\boldsymbol{\beta}}_{ols}$ 和残差 e_i；
2. 计算经过重缩放（rescaled）的残差：

$$u_i \equiv \left(\frac{n}{n-K}\right)^{1/2} e_i; \tag{3.36}$$

3. 利用如下式子生成 Bootstrap 观测：

$$y_i^* = \mathbf{x}_i'\hat{\boldsymbol{\beta}}_{ols} + u_i^*, \quad u_i^* \sim \text{EDF}\,(u_i),$$

其中，u_i^* 是从 u_i 的经验分布（EDF）中随机抽取的，这等价于对 $\{u_1, \ldots, u_n\}$ 进行有放回的抽样。

关于残差 Bootstrap 的小样本性质可以参考 MacKinnon (2006) 中提供的统计模拟结果。

Wild Bootstrap

考虑如下线性回归模型：

$$y_i = \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i, \quad \mathbb{E}(\varepsilon_i \mid \mathbf{x}_i) = 0.$$

非参数 Bootstrap 从原始样本中成对地[14]抽取 $\{y_i^*, \mathbf{x}_i^*\}$，这种抽样方式没有利用零条件期望假设，因此非参数 Bootstrap 样本分布不满足零条件期望假设。此外，非参数 Bootstrap 没有对 $\boldsymbol{\beta}$ 施加任何约束；并且，与残差 Bootstrap 和 Wild Bootstrap 不同的是，非参数 Bootstrap 并不以真实的数据矩阵 $\mathbf{X}$ 为条件。残差 Bootstrap 通过固定 $\mathbf{x}_i$ 并从重缩放的残差的经验分布中随机抽取 u_i^*，使得 Bootstrap 样本的回归误差项 u_i^* 独立于 $\mathbf{x}_i$，但是独立性假设强于零条件期望假设。此外，残差 Bootstrap 要求 ε_i 独立同分布。

Wu (1986) 提出的 Wild Bootstrap 施加了零条件期望假设，并且允许误差项存在异方差。获取 Wild Bootstrap 样本的一般步骤为：

1. 计算最小二乘估计量 $\hat{\boldsymbol{\beta}}_{ols}$ 和残差 e_i；
2. 利用如下式子生成 Bootstrap 观测：

$$y_i^* = \mathbf{x}_i'\hat{\boldsymbol{\beta}}_{ols} + e_i \cdot \xi_i^*, \quad \xi_i^* \sim i.i.d.\ (0,1),$$

其中，辅助随机变量 ξ_i^* 的分布有如下两种选择：

- Rademacher 随机数：

$$\xi_i^* = \begin{cases} 1 & \text{概率为 } \frac{1}{2} \\ -1 & \text{概率为 } \frac{1}{2}. \end{cases}$$

- Mammen (1993) 两点分布：

$$\xi_i^* = \begin{cases} \frac{(1-\sqrt{5})}{2} & \text{概率为 } \frac{(\sqrt{5}+1)}{2\sqrt{5}} \\ \frac{(\sqrt{5}+1)}{2} & \text{概率为 } \frac{(\sqrt{5}-1)}{2\sqrt{5}}. \end{cases}$$

Davidson et al. (2008) 中的统计模拟结果显示当 ε_i 具有对称的分布或近似对称的分布时，采用 Rademacher 随机数的 Wild Bootstrap 假设检验具有较好的小样本性质。若 ε_i 的分布不对称，建议使用 Mammen (1993) 两点分布。

注 在第 2 步中，常对 e_i 做如下的变换：

$$f(e_i) = \frac{e_i}{(1-h_{ii})^{1/2}}, \tag{3.37}$$

其中，h_{ii} 是第 i 个观测对应的杠杆值，见定义 3.1。

Bootstrap t 检验：几种方法的比较

为了比较不同的 Bootstrap 抽样方式对 Bootstrap 假设检验的小样本性质的影响，我们使用 R 语言编写几种 Bootstrap 的程序，通过统计模拟复现 MacKinnon (2006) 中图 1 的结果。

[14] 非参数 Bootstrap 也被称为 pairs Bootstrap。

MacKinnon (2006) 使用了如下的 DGP：

$$y_t = \beta_1 + \beta_2 y_{t-1} + \varepsilon_t, \quad \varepsilon_t \sim i.i.d.\, \mathcal{N}(0, \sigma^2),$$

其中，$\beta_1 = 1$，$\sigma^2 = 1$。考虑单个回归系数的假设检验问题：原假设为 $\mathbb{H}_0 : \beta_2 = 0.9$；备择假设为 $\mathbb{H}_1 : \beta_2 \neq 0.9$。为了检验 $\mathbb{H}_0$，构造 t 统计量：

$$t = \frac{\hat{\beta}_{2,ols} - 0.9}{s(\hat{\beta}_{2,ols})}.$$

在古典模型假设成立的情况下，t 服从自由度为 $n-2$ 的 t 分布。然而，上述 DGP 不满足严格外生性，$\hat{\beta}_{2,ols}$ 是有偏的，t 在小样本下的真实分布并非是 t 分布。这里选用 $\beta_2 = 0.9$ 作为原假设是因为此时的 DGP 虽然是平稳的但是它接近单位根过程，标准正态分布不能很好地近似 t 统计量在小样本下的真实分布。上述数据生成过程的 R 代码如下所示：

```
#定义数据生成过程DGP
DGP <- function(N){
  data = arima.sim(n = N, list(ar = 0.9)) + 10
  return(data)
}
```

我们比较如下五种统计推断方法的小样本性质：

1. 基于 t 分布的统计推断；
2. 残差 Bootstrap，估计模型系数时使用有约束的回归，即，施加了假设 $\mathbb{H}_0$ 的回归，并对有约束回归的残差进行式 (3.36) 的重缩放，记为 RR Bootstrap；
3. 残差 Bootstrap，估计模型系数时使用无约束的回归，并对无约束回归的残差进行式 (3.36) 的重缩放，记为 UR Bootstrap；
4. Pairs Bootstrap；
5. Wild Bootstrap，使用 Rademacher 随机数，并对残差进行式 (3.37) 的变换。

在比较小样本性质时，我们选取的样本量分别为 10、14、20、28、40、56、80、113、160、226、320、452、640、905 和 1280。针对每个样本量，重复实验的次数是 10000，每次重复时取 $B = 399$。初学计量的读者可能会对通过统计模拟评估检验方法小样本性质时涉及的重复实验和 Bootstrap 重抽样产生混淆，为了更好地理解它们之间的关系，我们利用流程图 3.2 说明统计模拟的过程。

从图中可以看出，给定样本量 n 之后，我们利用计算机生成 10^4 个不同的样本，对每一个样本做 399 次的 Bootstrap 得到检验的 Bootstrap p 值，这样可以得到 10^4 个 Bootstrap p 值：$\{p_1, \ldots, p_{10000}\}$，将这些 Bootstrap p 值和显著性水平 α 进行比较，计算其小于 α 的频率。对于一个理想的统计推断方法，此频率应该接近 α。

我们以 RR Bootstrap 为例说明如何编程实现 Bootstrap p 值的计算，以及如何计算 10^4 次重复实验的拒绝率，本小节其余四种统计推断方法以及完整的 R 程序可在本书网站下载，文件名为 boottest.R。由于统计模拟比较耗时，我们使用 parallel 包提供的并行计算功能。

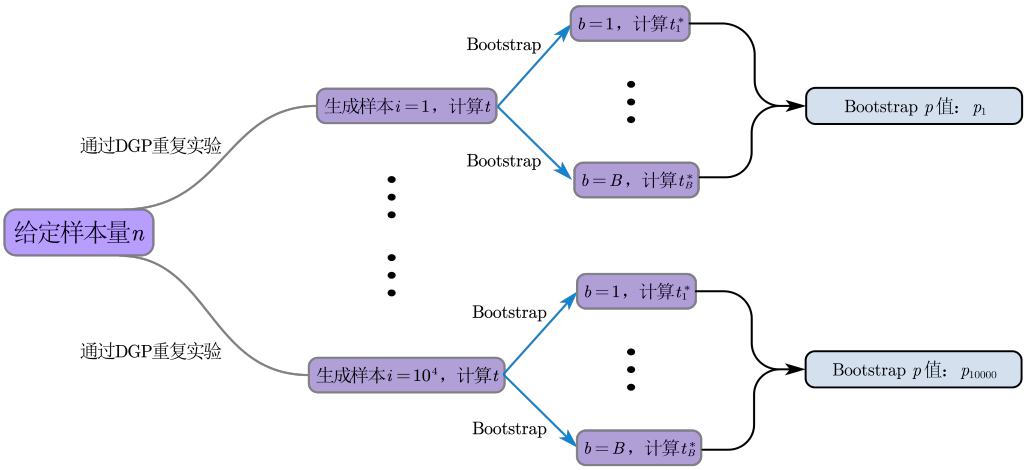

图 3.2: Bootstrap t 检验小样本性质评估的统计模拟流程图

```
#计算最小二乘估计量和t统计量
ols <- function(data){
  n = length(data)
  x = data[1:(n-1)]
  y = data[2:n]
  b = coef(lm(y~x))                                  #提取回归系数
  se = coef(summary(lm(y~x)))[,"Std. Error"][2]      #斜率估计量的标准误
  t = (b[2]-0.9)/se                                  #计算t统计量
  return(list(coef=b,std_error=se,test=t))
}
#RR Bootstrap
library(tidyverse)
RR_bootstrap <- function(data){
  n = length(data)
  xx = data[1:(n-1)]
  y = data[2:n]
  b0 = mean(y - xx*0.9)                              #有约束回归
  e = y - b0 - xx*0.9                                #计算有约束的残差
  u = sqrt((n-1)/(n-1-2))*e                          #式(3.36)
  data_star = rep(0,n)
  t_RR = lapply(
    1:B,
    function(x){
      data_star[1] = xx[1]
      for (i in 2:n){
        data_star[i] = b0 + 0.9*data_star[i-1] +
```

```
            sample(u,size=1,replace=TRUE)
      }
      data_star                                    #Bootstrap样本
    } %>%
      ols(.) %>%
      .["test"]                                    #计算t统计量
    )
  t_RR = unlist(t_RR)
  t = ols(data)$test
  p_RR = sum(abs(t_RR)>abs(t))/B
  return(p_RR)
}
library(parallel)                                  #加载并行计算包
clnum <- detectCores()                             #CPU核数
cl <- makeCluster(getOption("cl.cores", clnum))
#定义计算10000次模拟拒绝率的函数
reject_fre <- function(N){
  data = replicate(10000, DGP(N))
  p = parApply(cl,data,2,RR_bootstrap)
  return(mean(p<0.05))
}
clusterExport(cl, c("B","reject_fre","ols","RR_bootstrap","%>%"))
sample_size <- c(10,14,20,28,40,56,80,113,160,226,320,452,640,905,1280)
rej_fre = sapply(sample_size,reject_fre)
```

将不同的样本量做为横轴，各个样本量对应的 Bootstrap p 值小于 $\alpha = 0.05$ 的频率做为纵轴，我们得到图 3.3 所示的各个统计推断方法的小样本表现，从图中可以看出，有约束的残差 Bootstrap 和 Wild Bootstrap 具有良好的小样本性质，对于样本量的各种不同取值，两种检验方法的真实显著性水平都十分接近名义水平 0.05。

注 图 3.3 汇报的 Wild Bootstrap 结果优于 MacKinnon (2006) 中图 1 汇报的 Wild Bootstrap 结果，可能的原因是原文中的 Wild Bootstrap 采用的是无约束 Bootstrap，读者可以编程验证这一猜测。

聚类数据的 Bootstrap

尽管相关的理论文献较少，Bootstrap 方法依然可以用于聚类数据。本小节的讨论主要参考了 Cameron et al. (2008)。

令 $\mathbf{y}_g = (y_{1g}, \ldots, y_{n_g g})'$ 表示 $n_g \times 1$ 维第 g 个聚类中的所有被解释变量观测值构成的向量；$\mathbf{X}_g = (\mathbf{x}_{1g}, \ldots, \mathbf{x}_{n_g g})'$ 表示 $n_g \times K$ 维第 g 个聚类中的所有解释变量观测值构成

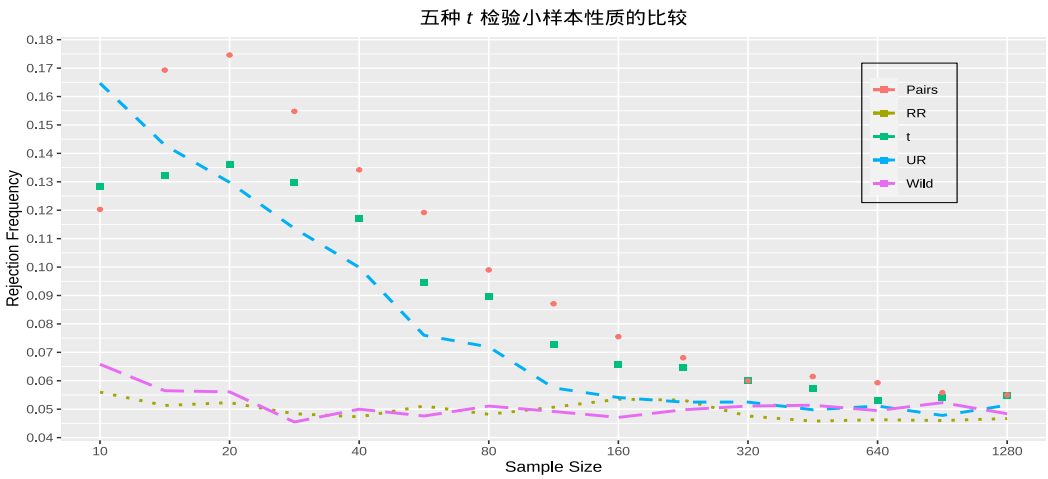

图 3.3: 五种 t 检验的小样本性质比较

的数据矩阵。聚类层面的线性回归方程可以表示为：

$$\mathbf{y}_g = \mathbf{X}_g\boldsymbol{\beta} + \boldsymbol{\varepsilon}_g, \quad g = 1, \ldots, G,$$

其中，$\boldsymbol{\varepsilon}_g = (\varepsilon_{1g}, \ldots, \varepsilon_{n_g g})'$ 是第 g 个聚类中的 $n_g \times 1$ 维的误差向量。原始样本可以划分为 G 个成对的 $\{\mathbf{y}_g, \mathbf{X}_g\}$。聚类的 pairs Bootstrap 对这 G 个 $\{\mathbf{y}_g, \mathbf{X}_g\}$ 进行有放回的 G 次抽样，从而得到 Bootstrap 样本并计算感兴趣的参数估计值或检验统计量的取值。将此过程重复进行 B 次，我们便可以计算 Bootstrap 标准误、构造各种 Bootstrap 置信区间或者计算 Bootstrap 假设检验的 p 值。值得一提的是，若要构造 BC$_a$ 百分位数区间，估计加速参数 a 时要使用“删除聚类”的 Jackknife。

获取聚类的 Wild Bootstrap 样本的一般步骤为：

1. 计算最小二乘估计量 $\hat{\boldsymbol{\beta}}_{ols}$ 和残差 e_i；
2. 利用如下式子生成 Bootstrap 观测：

$$\begin{aligned} \mathbf{y}_g^* &= \mathbf{X}_g\hat{\boldsymbol{\beta}}_{ols} + \boldsymbol{e}_g^* \\ \boldsymbol{e}_g^* &= \boldsymbol{e}_g \cdot \xi_g^* \end{aligned}$$

其中，$\boldsymbol{e}_g = (e_{1g}, \ldots, e_{n_g g})'$，表示第 g 个聚类的回归残差构成的列向量；ξ_g^* 是辅助标量随机变量，可以是 Rademacher 随机数，或者 Mammen (1993) 两点分布。

在 Stata 中，聚类 Bootstrap 标准误和置信区间可以通过在回归命令中设置选项 cluster(id) vce(bootstrap, reps(#)) 得到。例如：

```
//接例3.6的程序
. reg testscore tracking girl agetest, cluster(schoolid) vce(bootstrap,
    reps(500))
```

3.7 总结

球形扰动假设在建立最小二乘估计量的小样本性质时扮演着重要的角色。本章首先讨论了如何检验异方差，这是传统计量教材中的重要内容。若同方差的原假设成立，直接利用异方差条件下的结论进行统计推断会损失效率；然而，若是错误地接受了原假设，基于同方差条件下的理论进行的统计推断不具有稳健性。相比于效率，统计推断的稳健性往往更加重要。在传统的计量教材中，异方差被看作是一个“问题”，需要检验它是否存在，若存在该如何处理。然而，正如我们在第 2.1 节所讨论的那样，异方差是常态，同方差才是特例！被广泛使用的本科计量教材 Stock et al. (2020) 没有涉及任何关于异方差检验的讨论。因此，现在主流的计量实证分析中很少会出现异方差检验的使用，而是直接使用异方差稳健标准误。在 Stata 中，这种做法体现在 reg 命令后通常带着“, r”的选项。通过本章的学习，我们知道选项“, vce (hc2)”或“, vce (hc3)”可能是更好的选择。

类似地，当误差项存在自相关时，我们可以使用异方差自相关稳健标准误进行统计推断。时间序列模型是计量经济学研究的主要对象之一，时间序列数据往往会存在自相关和（或）异方差，在计算标准误时进行 Newey-West 调整显得尤为重要。Bali et al. (2016) 将 Newey-West 调整视为实证资产定价的三个预备知识之一，其重要性可见一斑。

自从二十世纪八十年代以来，聚类标准误在经济和金融的实证分析中被广泛地使用。本章以聚类数据为基础，讨论了使用聚类标准误的必要性。本书还将在因果推断（第 **??** 章）和静态面板数据模型（第 5 章）等章节进一步讨论聚类标准误的使用。近年来，关于聚类标准误的理论文献开始逐渐增多，感兴趣的读者可以参考 MacKinnon et al. (2022a) 和 Abadie et al. (2022)。值得一提的是，Abadie et al. (2022) 指出了使用聚类标准误的三个误区：（1）使用聚类标准误是因为同一聚类内部的个体的误差项存在相关性；（2）如果聚类标准误与异方差稳健标准误差异较大，我们就应该使用聚类标准误；（3）研究者只有两种选择：要么使用聚类标准误，要么使用异方差稳健标准误。

本章的最后介绍了 Jackknife 和 Bootstrap 两种重抽样方法。Jackknife 是计算标准误的既简单又灵活的统计方法。在通常情况下，Jackknife 方差的估计结果与渐近理论提供的渐近方差非常接近。Bootstrap 可以用来估计标准误、构造置信区间和假设检验。在估计回归系数的非线性函数的标准误时，截尾的 Bootstrap 标准误具有较好的稳定性。此外，不同的 Bootstrap 抽样方式会对假设检验的结果带来较大的影响。在异方差条件下，Wild Bootstrap 是较好的选择。在实证分析时，我们往往需要结合具体问题设计适合的 Bootstrap 方法（例如，Huang et al. (2015)）。

在实证分析中，研究者通常假设样本是从总体中随机抽取的一组观测，参数估计的标准误体现了抽样的变异（sampling variation）。然而，随着科技的发展和数据收集方式的改变，在大数据时代研究者可能拥有整个总体的数据（例如：京东的所有消费者），此时，我们该如何解释标准误？Abadie et al. (2020) 给出了一个答案，供读者参考。

附录 1：定理 3.4 的证明

首先，由最小二乘估计量的一致性和渐近正态性，我们有：

$$
\begin{aligned}
\sqrt{T}\hat{\gamma}_j &= \sqrt{T}\tilde{\gamma}_j - \frac{1}{T}\sum_{t=j+1}^{T}(\mathbf{x}_{t-j}\varepsilon_t + \mathbf{x}_t\varepsilon_{t-j})'\sqrt{T}(\hat{\boldsymbol{\beta}}_{ols} - \boldsymbol{\beta}) \\
&\quad + \sqrt{T}(\hat{\boldsymbol{\beta}}_{ols} - \boldsymbol{\beta})'\left(\frac{1}{T}\sum_{t=j+1}^{T}\mathbf{x}_t\mathbf{x}_{t-j}'\right)(\hat{\boldsymbol{\beta}}_{ols} - \boldsymbol{\beta}) \\
&\underset{a}{\sim} \sqrt{T}\tilde{\gamma}_j - \frac{1}{T}\sum_{t=j+1}^{T}(\mathbf{x}_{t-j}\varepsilon_t + \mathbf{x}_t\varepsilon_{t-j})'\sqrt{T}(\hat{\boldsymbol{\beta}}_{ols} - \boldsymbol{\beta}) \\
&\underset{a}{\sim} \sqrt{T}\tilde{\gamma}_j - \mathbb{E}(\mathbf{x}_{t-j}\varepsilon_t + \mathbf{x}_t\varepsilon_{t-j})'\sqrt{T}(\hat{\boldsymbol{\beta}}_{ols} - \boldsymbol{\beta}) && \text{(由遍历定理 B.7)} \\
&= \sqrt{T}\tilde{\gamma}_j - \boldsymbol{\mu}_j'\sqrt{T}(\hat{\boldsymbol{\beta}}_{ols} - \boldsymbol{\beta}) && (\boldsymbol{\mu}_j \equiv \mathbb{E}(\mathbf{x}_t\varepsilon_{t-j})) \\
&= \sqrt{T}\tilde{\gamma}_j - \boldsymbol{\mu}_j'\sqrt{T}\mathbf{S}_{\mathbf{xx}}^{-1}\frac{1}{T}\sum_{t=1}^{T}\mathbf{x}_t\varepsilon_t && \text{(由式 (2.36))} \\
&= \sqrt{T}\frac{1}{T}\sum_{t=j+1}^{T}\varepsilon_t\varepsilon_{t-j} - \boldsymbol{\mu}_j'\sqrt{T}\mathbf{S}_{\mathbf{xx}}^{-1}\frac{1}{T}\sum_{t=1}^{T}\mathbf{x}_t\varepsilon_t \\
&\underset{a}{\sim} \sqrt{T}\frac{1}{T}\sum_{t=1}^{T}\varepsilon_t\varepsilon_{t-j} - \boldsymbol{\mu}_j'\sqrt{T}\boldsymbol{\Sigma}_{\mathbf{xx}}^{-1}\frac{1}{T}\sum_{t=1}^{T}\mathbf{x}_t\varepsilon_t \\
&= \mathbf{c}_j'\sqrt{T}\bar{\mathbf{g}}_j,
\end{aligned}
$$

其中，

$$
\mathbf{c}_j \equiv \begin{bmatrix} 1 \\ -\boldsymbol{\Sigma}_{\mathbf{xx}}^{-1}\boldsymbol{\mu}_j \end{bmatrix}, \quad \bar{\mathbf{g}}_j \equiv \frac{1}{T}\sum_{t=1}^{T}\mathbf{g}_{jt}, \quad \mathbf{g}_{jt} \equiv \begin{bmatrix} \varepsilon_{t-j}\varepsilon_t \\ \mathbf{x}_t\varepsilon_t \end{bmatrix}.
$$

令 $\hat{\boldsymbol{\gamma}} = (\hat{\gamma}_1, \ldots, \hat{\gamma}_p)'$，由上述推导可知：

$$
\sqrt{T}\hat{\boldsymbol{\gamma}} \underset{a}{\sim} \mathbf{C}'\sqrt{T}\bar{\mathbf{g}}, \tag{3.38}
$$

其中，

$$
\mathbf{C} \equiv \begin{bmatrix} \mathbf{c}_1 & & \\ & \ddots & \\ & & \mathbf{c}_p \end{bmatrix}, \quad \bar{\mathbf{g}} \equiv \frac{1}{T}\sum_{t=1}^{T}\mathbf{g}_t, \quad \mathbf{g}_t \equiv \begin{bmatrix} \mathbf{g}_{1t} \\ \vdots \\ \mathbf{g}_{pt} \end{bmatrix}.
$$

由 LIE 和式 (3.3)，可以证明 $\mathbf{g}_t$ 是鞅差序列。显然，$\mathbf{g}_t$ 也是遍历平稳的，因此，

$$
\sqrt{T}\bar{\mathbf{g}} \underset{d}{\to} \mathcal{N}(\mathbf{0}, \mathbb{E}(\mathbf{g}_t\mathbf{g}_t')). \tag{3.39}
$$

接下来我们计算 $\mathbb{E}(\mathbf{g}_t\mathbf{g}_t')$，它的第 (j,k) 位置上的分块为：

$$
\mathbb{E}(\mathbf{g}_{jt}\mathbf{g}_{kt}') = \begin{bmatrix} \mathbb{E}(\varepsilon_t^2\varepsilon_{t-j}\varepsilon_{t-k}) & \mathbb{E}(\mathbf{x}_t'\varepsilon_t^2\varepsilon_{t-j}) \\ \mathbb{E}(\mathbf{x}_t\varepsilon_t^2\varepsilon_{t-k}) & \mathbb{E}(\varepsilon_t^2\mathbf{x}_t\mathbf{x}_t') \end{bmatrix}.
$$

再根据 LIE 和假设 2.8*，容易验证：

$$\mathbb{E}(\mathbf{g}_{jt}\mathbf{g}_{kt}') = \begin{bmatrix} \sigma^4\delta_{jk} & \sigma^2\boldsymbol{\mu}_j' \\ \sigma^2\boldsymbol{\mu}_k & \sigma^2\boldsymbol{\Sigma}_{\mathbf{xx}} \end{bmatrix},$$

其中，当 $j=k$ 时，$\delta_{jk}=1$；当 $j\neq k$ 时，$\delta_{jk}=0$。由式 (3.38) 和 (3.39)，

$$\mathrm{Avar}(\hat{\boldsymbol{\gamma}}) = \mathbf{C}'\mathbb{E}(\mathbf{g}_t\mathbf{g}_t')\mathbf{C}.$$

此渐近协方差矩阵的第 (j,k) 位置上的元素为：

$$\mathbf{c}_j'\mathbb{E}(\mathbf{g}_{jt}\mathbf{g}_{kt}')\mathbf{c}_k = \sigma^4\cdot[\delta_{jk} - \boldsymbol{\mu}_j'\boldsymbol{\Sigma}_{\mathbf{xx}}^{-1}\boldsymbol{\mu}_k/\sigma^2].$$

因此，

$$\sqrt{T}\hat{\boldsymbol{\gamma}} \underset{d}{\to} \mathcal{N}(\mathbf{0}, \sigma^4\cdot(\mathbf{I}_p - \boldsymbol{\Phi})).$$

又因为 $\sqrt{T}\hat{\boldsymbol{\rho}} \underset{a}{\sim} \sqrt{T}\hat{\boldsymbol{\gamma}}/\sigma^2$，$\sqrt{T}\hat{\boldsymbol{\rho}}$ 的渐近分布与 $\sqrt{T}\hat{\boldsymbol{\gamma}}/\sigma^2$ 的渐近分布相同，于是，我们有：

$$\sqrt{T}\hat{\boldsymbol{\rho}} \underset{d}{\to} \mathcal{N}(\mathbf{0}, \mathbf{I}_p - \boldsymbol{\Phi}).$$

附录 2：杠杆值与残差

如果样本中的少数观测数据离大多数观测数据很远，它们可能会对最小二乘估计产生很大的影响，这些数据被称为异常值（outlier）。某个观测值对最小二乘估计的影响力可通过去掉此观测值后回归系数的变化来衡量。为了计算回归系数的变化，我们首先定义数据矩阵 $\mathbf{X}$ 的杠杆值（leverage value）。

定义 3.1. 杠杆值

数据矩阵 $\mathbf{X}$ 的杠杆值是由 $\mathbf{X}$ 定义的投影矩阵 $\mathbf{P}_{\mathbf{X}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ 的对角线元素。♣

杠杆值有 n 个，通常表示为 h_{ii}，$i=1,\ldots,n$。因为，

$$\mathbf{P}_{\mathbf{X}} = \begin{pmatrix} \mathbf{x}_1' \\ \mathbf{x}_2' \\ \vdots \\ \mathbf{x}_n' \end{pmatrix} (\mathbf{X}'\mathbf{X})^{-1} \begin{pmatrix} \mathbf{x}_1 & \mathbf{x}_2 & \cdots & \mathbf{x}_n \end{pmatrix},$$

所以，我们有：

$$h_{ii} = \mathbf{x}_i'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i.$$

由上式可知 h_{ii} 是观测向量 $\mathbf{x}_i$ 的标准化长度。杠杆值有如下性质：

性质 （杠杆值的性质）

- $0 \leqslant h_{ii} \leqslant 1$；
- $\sum_{i=1}^n h_{ii} = K$；
- 如果 $\mathbf{X}$ 中包含常数项，$h_{ii} \geqslant 1/n$。

证明 令 $\boldsymbol{\tau}_i$ 表示第 i 个位置是 1 其余位置是 0 的 $n\times 1$ 维向量，则有 $h_{ii} = \boldsymbol{\tau}_i'\mathbf{P}_{\mathbf{X}}\boldsymbol{\tau}_i$。由

Rayleigh 商定理 A.12，

$$\boldsymbol{\tau}_i'\boldsymbol{\tau}_i\lambda_{\min}(\mathbf{P_X}) \leqslant h_{ii} \leqslant \boldsymbol{\tau}_i'\boldsymbol{\tau}_i\lambda_{\max}(\mathbf{P_X}).$$

注意到 $\mathbf{P_X}$ 是对称幂等阵，其特征值只能是 0 和 1，因此，$0 \leqslant h_{ii} \leqslant 1$。同时，

$$\sum_{i=1}^{n} h_{ii} = \operatorname{tr}(\mathbf{P_X}) = K.$$

此外，令 $\boldsymbol{\iota}$ 表示所有元素都是 1 的 $n \times 1$ 维向量，我们有：

$$\frac{1}{n} = \|\mathbf{P}_{\boldsymbol{\iota}}\boldsymbol{\tau}_i\|^2 = \|\mathbf{P}_{\boldsymbol{\iota}}\mathbf{P_X}\boldsymbol{\tau}_i\|^2 \leqslant \|\mathbf{P_X}\boldsymbol{\tau}_i\|^2 = h_{ii},$$

其中，第二个等号用到了 $\mathbf{X}$ 中包含常数项使得 $\mathbf{P}_{\boldsymbol{\iota}}\mathbf{P_X} = \mathbf{P}_{\boldsymbol{\iota}}$；不等号成立使用的是勾股定理。 □

接下来我们利用杠杆值来研究第 i 个观测值的影响力。去掉第 i 个观测值后的最小二乘估计量为：

$$\hat{\boldsymbol{\beta}}_{(-i)} = \left(\sum_{j\neq i}\mathbf{x}_j\mathbf{x}_j'\right)^{-1}\left(\sum_{j\neq i}\mathbf{x}_j y_j\right) = \left(\mathbf{X}'\mathbf{X} - \mathbf{x}_i\mathbf{x}_i'\right)^{-1}\left(\mathbf{X}'\mathbf{y} - \mathbf{x}_i y_i\right) \tag{3.40}$$

“去一”估计量对应的拟合值和残差为：

$$\tilde{y}_i = \mathbf{x}_i'\hat{\boldsymbol{\beta}}_{(-i)}, \quad \tilde{e}_i = y_i - \tilde{y}_i.$$

我们有如下结论：

性质 （“去一”估计量和残差）“去一”最小二乘估计量及其对应的残差可以通过如下方式计算：

$$\hat{\boldsymbol{\beta}}_{(-i)} = \hat{\boldsymbol{\beta}}_{ols} - \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{x}_i\tilde{e}_i, \tag{3.41}$$

$$\tilde{e}_i = (1 - h_{ii})^{-1} e_i. \tag{3.42}$$

证明 将式 (3.40) 左右两端同时左乘 $\left(\mathbf{X}'\mathbf{X}\right)^{-1}\left(\mathbf{X}'\mathbf{X} - \mathbf{x}_i\mathbf{x}_i'\right)$ 可以得到：

$$\begin{aligned}\hat{\boldsymbol{\beta}}_{(-i)} - \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{x}_i\mathbf{x}_i'\hat{\boldsymbol{\beta}}_{(-i)} &= \left(\mathbf{X}'\mathbf{X}\right)^{-1}\left(\mathbf{X}'\mathbf{y} - \mathbf{x}_i y_i\right) \\ &= \hat{\boldsymbol{\beta}}_{ols} - \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{x}_i y_i.\end{aligned}$$

上式经整理可得：

$$\hat{\boldsymbol{\beta}}_{(-i)} = \hat{\boldsymbol{\beta}}_{ols} - \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{x}_i(y_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}}_{(-i)}) = \hat{\boldsymbol{\beta}}_{ols} - \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{x}_i\tilde{e}_i.$$

上式两端同时左乘 $\mathbf{x}_i'$ 可得：

$$\mathbf{x}_i'\hat{\boldsymbol{\beta}}_{(-i)} = \mathbf{x}_i'\hat{\boldsymbol{\beta}}_{ols} - h_{ii}\tilde{e}_i.$$

所以，

$$y_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}}_{(-i)} = y_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}}_{ols} + h_{ii}\tilde{e}_i,$$

即，$\tilde{e}_i = e_i + h_{ii}\tilde{e}_i$，经整理得：$\tilde{e}_i = (1 - h_{ii})^{-1} e_i$。 □

具有影响力的观测

由式 (3.41)，观测 i 的影响力可以通过如下系数估计的差来衡量：

$$\hat{\boldsymbol{\beta}}_{ols} - \hat{\boldsymbol{\beta}}_{(-i)} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{x}_i\tilde{e}_i.$$

我们也可以通过拟合值的变化衡量观测 i 的影响力：

$$\hat{y}_i - \tilde{y}_i = \mathbf{x}_i'\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{x}_i\tilde{e}_i = h_{ii}\tilde{e}_i.$$

我们称具有较大杠杆值的观测为高杠杆点。从上式可以看出，高杠杆点可能是具有影响力的观测，但是高杠杆点是否确实具有影响力还取决于“去一”残差 $\tilde{e}_i$。

在 Stata 中，命令“dfbeta”可以计算如下测算影响力的指标：

$$\text{DFBETA}_{ik} \equiv \frac{\hat{\beta}_{k,ols} - \hat{\beta}_{k,(-i)}}{s_{(-i)}^2\sqrt{(\mathbf{X}'\mathbf{X})_{kk}^{-1}}},$$

其中，DFBETA_{ik} 衡量去掉第 i 个观测对第 k 个解释变量系数的影响；$s_{(-i)}^2$ 为 σ^2 的“去一”估计量，具体地，

$$s_{(-i)}^2 = \frac{\sum_{j\neq i}\tilde{e}_j^2}{n-K-1}.$$

Belsley et al. (2004) 推荐使用 $2/\sqrt{n}$ 作为阈值，若 $|\text{DFBETA}_{ik}| > 2/\sqrt{n}$，观测 i 可被认为是具有影响力的。

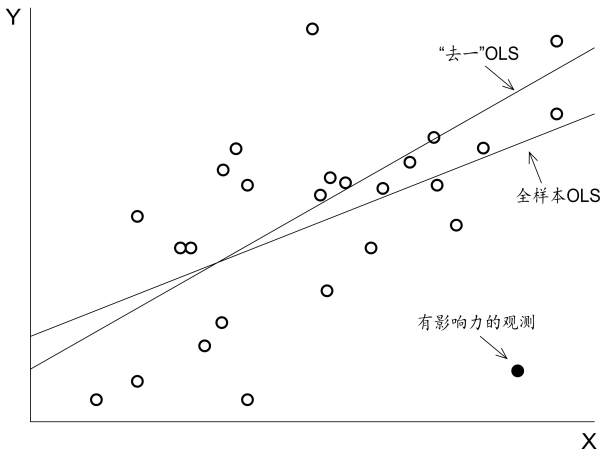

图 3.4: 具有影响力的观测对 OLS 的影响

残差

利用杠杆值，我们可以进一步分析残差 e_i 和“去一”残差 $\tilde{e}_i$ 的性质。令假设 2.1-2.4 成立，由式 (2.18)，

$$\begin{aligned}\mathbb{E}[\boldsymbol{e} \mid \mathbf{X}] &= \mathbb{E}[\mathbf{M}_{\mathbf{X}}\boldsymbol{\varepsilon} \mid \mathbf{X}] = \mathbf{0},\\ \text{Var}[\boldsymbol{e} \mid \mathbf{X}] &= \text{Var}[\mathbf{M}_{\mathbf{X}}\boldsymbol{\varepsilon} \mid \mathbf{X}] = \sigma^2\mathbf{M}_{\mathbf{X}}.\end{aligned}$$

因此，

$$\mathrm{Var}[e_i \mid \mathbf{X}] = \mathbb{E}(e_i^2 \mid \mathbf{X}) = (1 - h_{ii})\sigma^2. \tag{3.43}$$

由于 h_{ii} 是 $\mathbf{X}$ 的函数，所以，即使假设 2.4 成立，e_i 依然是异方差的。同时，e_i^2 是 σ^2 的有偏估计量。

类似地，对于“去一”残差向量 $\tilde{\boldsymbol{e}}$，由式 (3.42)，我们有：

$$\begin{aligned}
\tilde{\boldsymbol{e}} &= \mathbf{M}^*\mathbf{M}_{\mathbf{X}}\boldsymbol{\varepsilon}, \\
\mathbb{E}[\tilde{\boldsymbol{e}} \mid \mathbf{X}] &= \mathbb{E}[\mathbf{M}^*\mathbf{M}_{\mathbf{X}}\boldsymbol{\varepsilon} \mid \mathbf{X}] = \mathbf{0}, \\
\mathrm{Var}[\tilde{\boldsymbol{e}} \mid \mathbf{X}] &= \mathrm{Var}[\mathbf{M}^*\mathbf{M}_{\mathbf{X}}\boldsymbol{\varepsilon} \mid \mathbf{X}] = \sigma^2\mathbf{M}^*\mathbf{M}_{\mathbf{X}}\mathbf{M}^*
\end{aligned}$$

其中，$\mathbf{M}^* = \mathrm{diag}\left((1-h_{11})^{-1}, \ldots, (1-h_{nn})^{-1}\right)$。因此，

$$\begin{aligned}
\mathrm{Var}[\tilde{e}_i \mid \mathbf{X}] &= \mathbb{E}(\tilde{e}_i^2 \mid \mathbf{X}) = (1-h_{ii})^{-1}(1-h_{ii})(1-h_{ii})^{-1}\sigma^2 \\
&= (1-h_{ii})^{-1}\sigma^2.
\end{aligned}$$

若令 $\bar{e}_i = (1-h_{ii})^{-1/2}e_i$，容易验证 $\mathrm{Var}[\bar{e}_i \mid \mathbf{X}] = \sigma^2$。我们将 $\bar{e}_i$ 称为标准化残差。

再谈 σ^2 的估计

注意到 $\sigma^2 = \mathbb{E}(\varepsilon_i^2)$，因此估计 σ^2 的最直接方式是用 ε_i^2 的样本均值。然而，ε_i 不可观测，我们采用残差平方的样本均值：

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} e_i^2.$$

但是，由式 (3.43)，

$$\mathbb{E}[\hat{\sigma}^2 \mid \mathbf{X}] = \frac{1}{n}\sum_{i=1}^{n}(1-h_{ii})\sigma^2 = \frac{n-K}{n}\sigma^2, \tag{3.44}$$

其中，最后一个等号成立使用了杠杆值的和等于 K。所以，$\hat{\sigma}^2$ 是关于 σ^2 的有偏估计量。我们在定理 2.4 中考虑的估计量 s^2 通过在 $\hat{\sigma}^2$ 上乘以 $n/(n-K)$ 对偏误进行修正：

$$s^2 = \frac{\boldsymbol{e}'\boldsymbol{e}}{n-K} = \left(\frac{n}{n-K}\right)\frac{1}{n}\sum_{i=1}^{n} e_i^2.$$

但是，这不是唯一得到 σ^2 的无偏估计量的修正方式，例如：我们可以使用标准化残差平方的样本均值估计 σ^2：

$$\bar{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\bar{e}_i^2 = \frac{1}{n}\sum_{i=1}^{n}(1-h_{ii})^{-1}e_i^2.$$

容易验证 $\mathbb{E}(\bar{\sigma}^2 \mid \mathbf{X}) = \sigma^2$。

第三章习题

1. 证明：定理 3.2 中的 $\boldsymbol{g}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\boldsymbol{g} - n$ 等于 g_i 对 $\mathbf{z}_i$ 回归的回归平方和。
2. 证明：定理 3.3 的结论成立。
3. 证明：若 $\sqrt{T}\tilde{\gamma}_l - \sqrt{T}\hat{\gamma}_l \underset{p}{\to} 0$，则 $\sqrt{T}\tilde{\rho}_l - \sqrt{T}\hat{\rho}_l \underset{p}{\to} 0$。

4. 证明：在定理 3.4 的假设条件下，修正的 Box-Pierce Q 检验统计量与 $p \cdot F$ 渐近等价，其中，F 为检验辅助回归 (3.5) 中 $e_{t-1}, \ldots, e_{t-p}$ 系数联合显著性的 F 统计量。（选做）
5. 证明：式 (3.9) 成立。
6. 证明：式 (3.10) 和式 (3.11) 成立。
7. 证明：标准化残差 $\bar{e}_i$ 满足 $\mathrm{Var}[\bar{e}_i \mid \mathbf{X}] = \sigma^2$。
8. 证明：式 (3.21) 成立。
9. 计算：令 y^* 是从样本 $\{y_1, \ldots, y_n\}$ 中随机抽取的一个观测，F_n 是此样本的经验分布函数，求：$\mathbb{E}[y^* \mid F_n]$ 和 $\mathrm{Var}[y^* \mid F_n]$。
10. 计算：令 $\{y_1^*, \ldots, y_n^*\}$ 是对样本 $\{y_1, \ldots, y_n\}$ 进行有放回的抽样得到的 Bootstrap 样本，其样本均值为 $\bar{y}^* = n^{-1} \sum_{i=1}^n y_i^*$，求：$\mathbb{E}[\bar{y}^* \mid F_n]$ 和 $\mathrm{Var}[\bar{y}^* \mid F_n]$。
11. 编程：将图 3.3 中的 Wild Bootstrap 改为无约束的 Wild Bootstrap，通过统计模拟评估其小样本性质。